SANDRA CRISTINA DIAS NUNES

INCIDÊNCIAS - MODELO LOGIT E MEDIDAS APROXIMADAS DE IMPACTOS AMBIENTAIS

2 Volumes (Vol.I)

LISBOA

2006

SANDRA CRISTINA DIAS NUNES

INCIDÊNCIAS - MODELO LOGIT E MEDIDAS APROXIMADAS DE IMPACTOS AMBIENTAIS

2 Volumes (Vol.I)

Dissertação apresentada para obtenção do Grau de Doutor em Matemática na especialidade de Estatística pela Universidade Nova, Faculdade de Ciências e Tecnologia.

LISBOA

2006

”nº de arquivo”

”copyright”

"*O segredo de progredir é começar. O segredo é dividir as tarefas árduas e complicadas em tarefas pequenas e fáceis de executar, e depois começar pela primeira.*" (Mark Twain)

"O *êxito não se consegue só com qualidades especiais. É sobretudo um trabalho de constância, de método e de organização.*" (J.P. Sergent)

Aos meus Pais

Agradecimentos

Esta tese só se tornou possível pela solidariedade e contribuição de diversas pessoas. Algumas, como diria, o poeta Brecht, foram imprescindíveis; outros, fundamentais; e o mais singelo gesto de ajuda também foi importante. Não mencionarei todos, pois a lista seria grande. Mas não poderia deixar de assinalar os quais considero imprescindíveis.

Em primeiro lugar agradeço aos meus orientadores o Professor João Tiago Mexia e o Professor Christoph Minder.

Ao Professor Tiago Mexia, a quem devo a orientação geral dos trabalhos, quero agradecer o constante empenho pessoal manifestado, o encorajamento e apoio dado ao longo de todo o trabalho. Com certeza não teria conseguido sem a sua orientação.

Ao Professor Christoph Minder que, apesar de se encontrar numa situação extremamente difícil, nunca deixou de se empenhar neste trabalho, tendo sido fundamentais as suas ideias e sugestões.

Aos dois o meu *muito obrigado*.

Agradeço a todos os amigos e colegas por todo o apoio e interesse manifestados. Em particular, quero agradecer à Ana Isabel Matos, pelos valiosos comentários e sugestões.

Finalmente, porque os últimos são os primeiros, agradeço aos meus Pais, a quem dedico esta tese, pelo apoio incondicional e pela paciência ilimitada.

Resumo

O objectivo central deste trabalho é avaliar o efeito de se considerarem medições aproximadas das exposições na incidência de doenças resultantes de impactos ambientais.

Dado existirem quase exclusivamente estatísticas de incidência das doenças, desenvolvemos modelos que, a partir dessas estatísticas, permitam abordar o nosso problema. Em particular, mostrámos que, quando se utilizam *modelos logit*, o uso de medidas aproximadas dos impactos ambientais leva a uma distorção, por defeito, dos coeficientes angulares das rectas ajustadas.

Estudámos ainda limites para essa distorção, utilizando para tal a aproximação de Edgeworth.

Os nossos resultados permitiram-nos também esquematizar diversos cenários para delineamento de trabalhos de campo, indispensáveis ao aprofundamento do nosso objectivo.

Abstract

The main goal of this essay is to evaluate the effect of considering approximate measurements of the exposure on the incidence of diseases that result from environmental impacts.

Since the available statistics are, almost exclusively, on diseases incidence, we developed models that allow us to approach our problem starting from those statistics. In fact, we showed that when *logit models* are used the use of approximate measures of environmental impacts leads to a negative bias for adjusted slopes.

We also used the Edgeworth expansions to deduct bounds for that bias.

Furthermore, our results allowed us to schematize several scenarios for "field work". This "field work"is essential to the fulfilling of our goal.

Lista de Símbolos

As matrizes e os vectores são ambos representados a "**bold**". É importante referir que as matrizes são sempre representadas por letras maiúsculas, enquanto que os vectores surgem ao longo deste texto em ambos os formatos, representados quer por letras minúsculas quer por maiúsculas. Tem-se assim:

$\boldsymbol{X}_{m\times n}$ representa uma matriz com m linhas e n colunas; no entanto, sempre que o texto o permita, escrever-se-á apenas $\boldsymbol{X}$;

$\boldsymbol{y}^m$ **ou** $\boldsymbol{Y}^m$ representa um vector com m componentes, isto é, uma matriz com m linhas e 1 coluna; tal como no caso anterior, sempre que não haja lugar a confusões, escrever-se-á simplesmente $\boldsymbol{y}$ ou $\boldsymbol{Y}$;

y_i representa a i-ésima componente do vector $\boldsymbol{y}$;

x_{ij} representa o elemento da matriz $\boldsymbol{X}$ que se encontra na linha i e coluna j;

$\boldsymbol{X}^T$ representa a transposta de $\boldsymbol{X}$;

$\boldsymbol{y}^n_\nabla$ representa a projecção ortogonal de $\boldsymbol{y}^n$ sobre ∇;

$\Omega(\boldsymbol{y}^n)^\perp$ representa o complemento ortogonal de $\Omega(\boldsymbol{y}^n)$;

$D = diag(\boldsymbol{p}^n)$ representa uma matriz diagonal cujos elementos da diagonal principal correspondem às componentes do vector $\boldsymbol{p}^n$;

$\Re(\boldsymbol{X})$ representa o espaço imagem de $\boldsymbol{X}$;

$\boldsymbol{X}^{-1}$ representa a matriz inversa da matriz $\boldsymbol{X}$;

$\boldsymbol{X}^-$ representa uma inversa generalizada da matriz $\boldsymbol{X}$;

$\boldsymbol{X}^\dagger$ representa a matriz inversa generalizada de Moore-Penrose da matriz $\boldsymbol{X}$;

$car(\boldsymbol{X})$ representa a característica da matriz $\boldsymbol{X}$;

$\#(C)$ representa o cardinal do conjunto C;

$\|\boldsymbol{y}^n\|$ representa a norma do vector $\boldsymbol{y}^n$;

χ^2_s representa uma variável aleatória com distribuição qui-quadrado central, com s graus de liberdade;

$N(\mu, \sigma^2)$ representa uma variável aleatória com distribuição normal, com valor esperado μ e variância σ^2;

$\overset{\bullet}{\sim}$ representa "aproximadamente distribuído";

$\approx$ representa "aproximadamente igual a";

$\ll$ representa "significativamente inferior a";

$\gg$ representa "significativamente superior a";

$\otimes$ representa o produto de Kronecker de matrizes;

δ_{kt} representa o símbolo de Kronecker;

R^2 representa o coeficiente de determinação;

$\mathbb{R}$ representa o conjunto dos números reais.

Lista de Abreviaturas

ln representa o logaritmo natural ou neperiano;

UMP representa "Uniformemente Mais Potente";

VFM representa "Variance Free Models";

TB representa Tuberculose;

WHO representa World Health Organization;

GE representa Grupo Etário;

GE A representa Grupo Etário A;

GE B representa Grupo Etário B;

GE C representa Grupo Etário C;

GEAF ou AF representa o conjunto de indivíduos do sexo feminino do grupo etário A;

GEAM ou AM representa o conjunto de indivíduos do sexo masculino do grupo etário A;

GEBF ou BF representa o conjunto de indivíduos do sexo feminino do grupo etário B;

GEBM ou BM representa o conjunto de indivíduos do sexo masculino do grupo etário B;

GECF ou CF representa o conjunto de indivíduos do sexo feminino do grupo etário C;

GECM ou CM representa o conjunto de indivíduos do sexo masculino do grupo etário C;

ANOVA representa Analysis of Variance;

OV representa Origem da Variação;

SQ representa Soma de Quadrados;

GL representa Graus de Liberdade;

QM representa Média de Quadrados;

$\mathcal{F}$ representa Valor da Estatística $\mathcal{F}$;

ds representa a diferença significativa resultante do teste de correlação múltipla de Scheffé.

Siglas dos Países

Albania = ALB

Alemanha = GER

Arménia = ARM

Áustria = AUT

Azerbaijão = AZE

Bélgica = BEL

Bósnia-Herzegovina = BIH

Bulgária = BUL

Cazaquistão = KAZ

Croácia = CRO

Dinamarca = DEN

Eslováquia = SVK

Eslovénia = SLO

Espanha = ESP

Estónia = EST

Finlândia = FIN

França = FRA

Geórgia = GEO

Grécia = GRE

Holanda = NED

Hungria = HUN

Irlanda = IRL

Islândia = ISL

Israel = ISR

Itália = ITA

Letónia = LAT

Lituânia = LTU

Luxemburgo = LUX

Macedónia = MAC

Malta = MLT

Moldávia (Rep.) = MDA

Noruega = NOR

Polónia = POL

Portugal = POR

Quirguistão = KGZ

Reino Unido = GBR

República Checa = CZE

Roménia = ROM

Rússia (Fed.) = RUS

Suécia = SWE

Suíça = SWI

Tajiquistão = TJK

Turquemenistão = TKM

Ucrânia = UKR

Usbequistão = UZB

Jugoslavia = YUG

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

É inegável que as metodologias para estudos que relacionam a saúde e o ambiente são necessariamente mais variadas e complexas do que nas outras áreas da saúde. A diversidade do conceito de ambiente aumenta o número de questões de interesse, que exigem diferentes formas de abordagem metodológicas; falamos de questões como a poluição química, o saneamento e a qualidade da água, a pobreza, a equidade, as condições psico-sociais, etc.

Na pesquisa em Saúde Ambiental existem ainda diversos campos em aberto, alguns praticamente inexplorados. Neste cenário, tem sido dada uma prioridade aos poluentes químicos ambientais, enquanto causadores de doenças. Esta é justificada pelo elevado número de substâncias químicas utilizadas nas diversas actividades económicas. Estes poluentes, que cada vez são em maior número e concentração, em especial nas grandes cidades, atingem a população provocando graves problemas de saúde. O crescimento deste problema de saúde pública contribuiu bastante para que a Epidemiologia passasse a ser fundamental neste campo da investigação.

De uma forma algo simplista, podemos dizer que a Epidemiologia Ambiental pretende descrever, analisar e, consequentemente, interferir na relação entre a exposição a poluentes ambientais e a ocorrência de efeitos adversos para a saúde das populações. Para além de todos os problemas que um estudo epidemiológico apresenta, no caso da poluição de origem química esses problemas são agravados pela elevada complexidade de tecnologias utilizadas na avaliação das exposições e efeitos, pela ausência de conhecimentos toxicológicos e, pela dificuldade em definir metodologias que permitam construir a população de referência. Esta diversidade de questões confere aos estudos em Saúde Ambiental uma maior complexidade metodológica. Como consequência lógica, os estudos epidemiológicos para a produção de conhecimento e acções de vigilância ambiental em saúde exigem como principal estratégia um trabalho integrado que contemple, além da participação da comunidade, a articulação de instituições de diversos sectores, o que acaba em grande maioria dos casos por dificultar a obtenção de dados, inviabilizando a aplicação prática das metodologias desenvolvidas.

Estas e outras ideias podem ser aprofundadas em Câmara e Tambellini (2003).

Inicialmente, pretendíamos estudar o efeito de medidas imprecisas das exposições nas incidências de doenças resultantes de impactos ambientais. Verificámos então a conveniência de utilizar o *modelo logit*, que genericamente pode ser apresentado na

forma

$$logit(p) = \beta_0 + \beta_1 z$$

onde p mede a taxa de incidência, z representa o impacto e β_0 e β_1 são os parâmetros do modelo.

No entanto, a falta de dados forçou-nos a aprofundar a teoria no sentido de poder utilizar tais modelos quando apenas se dispõe de taxas de incidência (Capítulo 3). Tal aprofundamento foi possível devido à utilização dos mínimos quadrados estruturados, obtendo-se assim estimadores para:

- os coeficientes do modelo, β_0 e β_1;
- as exposições.

Estes últimos estimadores podem ser considerados como medidas aproximadas das exposições.

No estudo teórico dos *modelos logit* que fizemos mostrámos que a utilização de tais medidas conduz a um enviesamento negativo na estimação de β_1 (capítulo 4). Ter-se-á assim uma perda de sensibilidade à variação da exposição. Será ainda de interpretar os estimadores obtidos para o β_1 como limites inferiores.

Numa segunda fase do nosso trabalho (capítulo 5), utilizámos os nossos resultados para delineamento de Estudos de Campo. Obtivemos para isso uma partição da variância dos estimadores de β_1 que nos permitiu considerar diversos cenários em que se propõe a introdução de estações de monitorização:

- **Primeiro Cenário** - as estações e as sub-populações a considerar estão previamente escolhidas;
- **Segundo Cenário** - as estações estão implantadas mas os limites das regiões a atribuir-lhes não estão definidos;
- **Terceiro Cenário** - apenas se tem uma ideia aproximada do número de estações a implantar;
- **Quarto Cenário** - o número de estações não é conhecido.

Para tornar este trabalho o mais auto-suficiente possível, começamos por, no Capítulo 2, apresentar um conjunto de resultados preliminares, com especial incidência no estudo da metodologia logit, que julgamos fundamentais para o desenvolvimento do trabalho apresentado nos capítulos seguintes.

No final do Capítulo 3, apresentamos ainda alguns exemplos de aplicação à incidência de Tuberculose e de SIDA.

Apresentam-se no último capítulo breves ideias para trabalhos futuros que permitam aperfeiçoar a metodologia desenvolvida nesta dissertação.

O trabalho é completado por dois apêndices, nos quais se apresentam cálculos mais pesados que permitiram a obtenção dos momentos do estimador de β_1 e também do bias, necessários para a construção dos intervalos de confiança. Convém salientar que o Apêndice B contém apenas as expressões finais de tais momentos, sendo que a globalidade dos cálculos pode ser consultada no Volume II desta dissertação.

Capítulo 2

Resultados Preliminares

Dedicamos o segundo capítulo deste texto aos resultados que julgamos fundamentais para o desenvolvimento desta dissertação. Na primeira e mais importante secção deste capítulo é estudado o *Modelo Logit* e o *teorema de Fisher*. Na segunda secção é abordada a temática das aproximações à Normal, com especial atenção às *expansões de Edgeworth*. As duas últimas secções deste capítulo são dedicadas à apresentação de algumas noções básicas que serão utilizadas nos capítulos seguintes.

2.1 Métodos Ligados à Normal

A distribuição Normal é, sem sombra de dúvida, um dos pilares da teoria estatística, sendo fundamental, para não dizer imprescindível, em qualquer estudo desta natureza. Assim, todos os métodos que de alguma forma se encontrem ligados a esta distribuição são importantes.

2.1.1 Modelo Logit

O Papel do Modelo Logit

Os modelos de regressão passaram a ser parte integrante de qualquer análise cujo objectivo seja descrever a relação existente entre a variável resposta e uma ou mais covariáveis. Não é, por isso, de espantar que afirmemos ser este o papel do *Modelo Logit* ou *Modelo de Regressão Logística*, sinónimos para o mesmo método. A *análise logit* é considerada como complemento natural da familiar regressão linear sempre que o resultado não se exprime através duma variável contínua, mas representa uma determinada situação que pode ou não verificar-se; por outras palavras, quando a variável resposta é discreta com duas determinações possíveis. Há mais de uma década que o *Modelo de Regressão Logística* se tornou, nas mais diversas áreas, o método standard quando estamos perante problemas desta natureza.

Podemos dizer, de uma forma algo simplista, que o que distingue o *Modelo de Regressão Logística* do Modelo de Regressão Linear é o facto da variável resposta no *Modelo Logit* ser uma variável binária ou dicotómica.

Num primeiro contacto, é normal ficarmos com a ideia de que estes dois métodos são bastante diferentes e que o *Modelo Logit* apresenta um aspecto bastante mais

complexo que a familiar regressão linear. Sob um olhar mais profundo chegamos à conclusão que, afinal, os dois métodos têm muito em comum.

Em ambos os modelos existe uma definida assimetria entre a variável independente ou covariável e a variável dependente ou variável resposta. Ambos foram concebidos para a análise de dados experimentais, dados para os quais o princípio da causalidade não se questiona. Neste contexto, o modelo de regressão linear oferece uma "imperfeita"mas "quase universal"estrutura para a análise empírica. Reconhecidamente, na maior parte das vezes não é mais do que uma aproximação simplificada de algo presumivelmente melhor. Servindo, dentro das suas limitações, para uma selecção empírica da evidência.

A *regressão logística* pode ser usada de forma em tudo semelhante quando o fenómeno em estudo reporta dados categorizados em duas classes.

Existem, tal como já referimos, diferenças entre os dois modelos. Ao contrário da regressão linear, o *Modelo Logit* permite uma interpretação economicista, através da maximização da utilidade em situações de escolha discreta. Esta propriedade faz com que os economistas confiram a este modelo um estatuto que vai além de uma conveniente técnica empírica.

Existe ainda uma subtil, mas importante, diferença entre os dois métodos, que reside no facto do modelo de regressão linear necessitar de um parâmetro "perturbador", enquanto que no *Modelo Logit* o carácter aleatório do resultado é parte integrante das especificações iniciais.

Conjuntamente com o modelo Probit, o *Modelo Logit* pertence à classe dos modelos probabilísticos que permitem determinar probabilidades discretas sobre um número limitado de resultados possíveis.

Tal como o modelo de regressão linear, também o *modelo Logit* permite os mais diversos e complexos tipos de extensões e variantes. Como exemplos, podemos citar o Modelo Logit Multinomial, ver MacFadden (1975), o Modelo "Nested"Logit, ver Ben-Akiva (1985), o Modelo Generalizado do Valor Extremo, ver McFadden (1978), etc.

Não é preciso um estudo muito aprofundado da bibliografia especializada para chegarmos à conclusão que um número significativo destas variações do *Modelo Logit* têm sido desenvolvidas num perfeito isolamento por investigadores de áreas como a Biologia, a Epidemiologia, a Medicina, a Econometria, a Economia, etc. Este desenvolvimento provocou o aumento da separação de paradigmas que partilham a mesma essência estatística mas que usam abordagens distintas, uma vez que se ocupam de dados de natureza diferente e procuram resultados distintos.

Cada disciplina apresenta, deste modo, um modelo "seu", que suporta a mesma técnica estatística mas com enormes diferenças na abordagem, na terminologia e mesmo na interpretação. Diferenças estas que, em muitos casos, são exageradas, atingindo dimensões ideológicas, criando imensos obstáculos na troca de informação entre investigadores de áreas diferentes.

O ideal será mesmo a leitura de variadas fontes peritas nesta temática e nas variantes provenientes das diversas áreas. Em seguida apresenta-se uma lista, que sabemos ser incompleta mas de certeza poderá fornecer conhecimentos importantes sobre esta matéria:

- na área "bio-assay"[1] sugerimos o clássico de Finney (1971), cuja primeira publicação data de 1947, e (como não podia deixar de ser) Berkson (1944);

- nas áreas da Epidemiologia e Medicina sugerimos a monografia de Hosmer e Lemeshow (2000) e o artigo sobre os estudos de caso-controlo de Breslow (1996);

- para uma visão puramente estatística não podemos deixar de ler Cox e Snell (1989), MacCullagh e Nelder (1989) e Agresti (1990), que é um tratado sobre variáveis categorizadas;

- para uma abordagem econométrica sugerimos Amemya (1981), Maddala (1983) e Gourieroux (2000);

- uma aplicação às Ciências Sociais pode ser estudada no artigo de Menard (1995);

- uma aplicação ao Marketing é apresentada em Franses e Paap (2001).

Um pouco de história

De forma sucinta, vamos descrever a origem da *função logística* e a sua evolução até aos dias de hoje. Esta breve resenha foi baseada no excelente artigo de J. S.Cramer (2002), intitulado "The origins of logistic regression".

Começamos por apresentar na Figura (2.1), a curva de forma sigmóide que representa a *função logística*

$$P(z) = \frac{e^z}{1+e^z}. \tag{2.1}$$

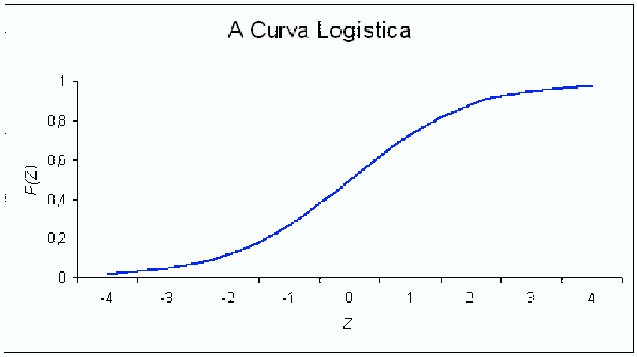

Figura 2.1: A Curva Logística.

[1]Bioassay é um processo que avalia a potência de um estímulo analisando os efeitos que causa num organismo biológico.

A função P tem um comportamento idêntico ao de uma função distribuição de uma densidade simétrica com ponto médio igual a zero. Assim, $P(z)$ é crescente e toma valores no intervalo $]0,1[$.

O significado da *função logística* varia de acordo com as variáveis envolvidas. Por exemplo, na aplicação do *modelo logit* à teoria "bio-assay", P representa a probabilidade de um resultado binário (a sobrevivência ou morte de um organismo) com $z = \alpha+\beta x$, onde x é a variável que representa a exposição (a um dado medicamento), α é a localização da curva no eixo dos xx e β a sua variação.

Originalmente a *função logística* (2.1) foi concebida para descrever o percurso da proporção P no decorrer do tempo t, com $z = \alpha+\beta t$, obtendo-se neste caso a "curva de crescimento"(2.1). Note-se que $P(t)$ aumenta monotonamente com o tempo t.

Se observarmos o gráfico da curva logística (Figura (2.1)), concluímos que para valores de P entre $(0,3)$ e $(0,7)$, a forma desta curva assemelha-se à da curva da função distribuição normal. Ambas as funções

$$P_t(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

e

$$P_n(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} (e^{(-1/2(u/\sigma)^2)})du$$

passam pelo ponto $(0,5)$ e são praticamente coincidentes através de um ajustamento apropriado dos parâmetros α e β. Convém referir que este facto não passa de uma total coincidência numérica, pois não parece existir qualquer relação entre as duas funções.

A *função logística* foi criada no século XIX com dois grandes objectivos - descrever o crescimento das populações e o percurso de reacções químicas autocatalíticas ou reacções em cadeia. Em ambos os casos, é definida uma função do tempo $W(t)$ e a sua taxa de crescimento

$$\dot{W}(t) = (dW(t))/dt.$$

Assumindo que $\dot{W}(t)$ é proporcional a $W(t)$ tem-se

$$\dot{W}(t) = \beta W(t), \tag{2.2}$$

isto é

$$\beta = \dot{W}(t)/W(t),$$

com β a constante que representa a taxa de crescimento. O que conduz a um crescimento exponencial

$$W(t) = Ae^{\beta t},$$

onde A é, por vezes, substituída pelo valor inicial $W(0)$.

Segundo Malthus (1798), a população humana, se não existissem interferências, tenderia a crescer segundo uma progressão geométrica. Este modelo chegou a ser considerado razoável quando aplicado a um país novo com uma população pequena.

Tal como muitos outros, também Alphonse Quetelet (1795-1874), um astrónomo belga que se tornou estatístico, estava convencido que a extrapolação indiscriminada

do crescimento exponencial conduziria a valores impossíveis. Quetelet pediu a um seu aluno, o matemático belga Pierre-François Verhulst (1804-1849), para estudar o problema.

Tal como Quetelet, Verhulst abordou o modelo adicionando à equação (2.2) um novo termo, que representa a resistência ao aumento do crescimento, obtendo a equação

$$\dot{W}(t) = \beta W(t) - \phi(W(t))$$

ou, de outra forma,

$$\dot{W}(t) = \beta W(t)(\Omega - W(t))$$

onde Ω denota o limite superior do nível de saturação de W, ou seja a sua assímptota quando $t \to \infty$.

Tomando-se $P(t) = W(t)/\Omega$, obtém-se

$$P(t) = \beta P(t)[1 - P(t)],$$

sendo a solução desta equação diferencial dada por

$$P(t) = \frac{e^{\alpha+\beta t}}{1 + e^{\alpha+\beta t}} \tag{2.3}$$

que Verhulst denominou de *função logística*.

A população $W(t)$ é, então, dada por

$$W(t) = \Omega \frac{e^{\alpha+\beta t}}{1 + e^{\alpha+\beta t}}. \tag{2.4}$$

Verhulst publicou as suas ideias relativamente a esta temática, em três artigos, entre 1838 e 1847. O primeiro foi publicado numa revista intitulada "Correspondance Mathématique et Physique", editada por Quetelet em 1838, sendo este o que contém a essência dos seus argumentos. Os dois artigos seguintes foram ambos publicados nos "Proceedings of Belgian Royal Academy". No segundo, publicado em 1845, Verhulst preocupa-se muito mais com a função e as suas propriedades, sendo neste artigo que pela primeira vez a denomina, sem grandes explicações, de *função logística*. Determina ainda os três parâmetros Ω, α e β da equação (2.4), fazendo com que a curva passe por três pontos observados.

Neste segundo artigo, utilizando a população belga nos anos de 1815, 1830 e 1845, Verhulst estabelece um limite para essa mesma população de 6,6 milhões e, num exercício similar, estabelece para a população francesa um limite de 40 milhões. Sendo conhecido, a essa data, que a população da Bélgica era de 10,2 milhões e a da França de 58,7 milhões, o método não foi muito bem recebido.

No terceiro e último artigo, publicado em 1847, Verhulst consegue corrigir o ajustamento e obtém uma estimativa para a população belga de 9,5 milhões, conferindo uma maior fiabilidade ao modelo.

No entanto, a descoberta da *função logística* por parte de Verhulst não foi bem recebida pelo seu professor, Quetelet, que segundo Vanpaemel (1987) não partilhava a maior parte das ideias do seu aluno. O trabalho de Verhulst foi citado com

aprovação apenas por Liagre (1852), seu colega na Academia Militar, onde ambos leccionavam.

Como um modelo que estuda o crescimento da população, a *função logística* foi redescoberta em 1920 por Pearl[2] e Reed[3]. Aparentemente, ambos desconheciam o trabalho de Verhuslt e obtiveram, individualmente, a *função logística*.

Quando este modelo foi ajustado aos censos dos Estados Unidos da América, novamente fazendo com que a curva passasse por três valores observados, o resultado para o período entre 1790 e 1910 foi aceitável. No entanto, a estimativa de Ω foi de 197 milhões que, mais uma vez, não se aproximava do valor real da população que era de 270 milhões.

Pearl e os seus colaboradores continuaram com o desenvolvimento do modelo e, durante os vinte anos seguintes, aplicaram a curva logística ao estudo das mais variadas espécies, desde a mosca da fruta até à população das colónias francesas no norte de África.

O trabalho de Verhulst foi redescoberto imediatamente após a publicação deste primeiro artigo de Pearl e Reed, em 1920. No entanto, só num artigo de 1923 estes autores mencionaram o trabalho do Matemático belga, chamando-lhe "o há muito esquecido".

A publicação de relevo que se seguiu deve-se a Yule e surgiu em 1925. Foi também este autor que fez reviver o nome "*logística*", que nem Liagre, Pearl ou Reed usaram nos seus artigos.

Foi preciso esperar até 1933 para ver publicado um artigo reconhecendo o trabalho de Verhulst. Este tributo foi prestado por um dos colaboradores de Pearl de nome Miner.

Como foi referenciado no início desta secção, a *função logística* foi concebida não só para estudar o crescimento das populações, mas também para estudar o curso de reacções catalíticas ou reacções em cadeia. Reed e Berkson[4], em 1929, apresentaram aplicações da *função logística*, com algumas variantes, a diversos processos desta natureza. Neste artigo é citado o trabalho de um professor alemão, Wilhem Ostwald, datado de 1883. O que vem mostrar que esta metodologia há muito tinha aplicações também nesta área.

Podemos, assim, concluir que a ideia fundamental da *função logística* é simples e eficiente, sendo até aos dias de hoje utilizada como um modelo de crescimento de populações e para estudar o curso de processos autocatalíticos de que são exemplo a introdução de novos produtos e tecnologias no mercado.

---

[2]Raymond Pearl (1879-1940) - a sua formação de base é a Biologia, tendo adquirido um forte treino como Estatístico quando passou um ano em Londres a trabalhar com Karl Pearson, tornando-se posteriormente num prodigioso investigador. Em 1920 foi nomeado Director do Departamento de Biometria e Estatística Vital da Universidade Johns Hopkins.

[3]Lowell J. Reed (1886-1966) - o "braço direito"de Pearl na Universidade. A formação de base é a Matemática. Fez uma carreira tranquila como Bioestatístico. Distinguiu-se como Professor e Administrador, tendo sido nomeado Presidente de Johns Hopkins em 1953.

[4]Joseph Berkson (1899-1982) - estudou Física em Columbia, tendo mudado para a Universidade Johns Hopkins onde se Doutorou em Estatística no ano de 1928. Permaneceu nesta universidade como assistente durante três anos onde trabalhou com Reed nas funções autocatalíticas. Mudou-se para a "Mayo Clinic"onde permaneceu até ao fim da sua carreira profissional como "Chief Statistician".

A evolução desta metodologia originou inúmeras variantes.

A invenção do modelo probit é usualmente atribuída a Gaddum[5] (1933) e Bliss[6] (1934-A, 1934-B), mas basta um breve olhar sobre a secção histórica do livro de Finney (1971) ou uma leitura do artigo de Gaddum para percebermos que tal não é totalmente correcto.

Através dessa leitura concluímos que este método surgiu pela mão do alemão Fechner (1801-1887), com uma aplicação na transformação de frequências em equivalentes desvios normais. Fechner verificou que a resposta de um ser humano a diferentes estímulos não é uniforme e foi o primeiro a transformar as diferenças observadas em equivalentes desvios normais.

A resenha histórica do livro de Finney refere um largo número de "redescobertas"desta temática que cobrem os setenta anos que medeiam o trabalho de Fechner (1860) e as publicações de Gaddum e Bliss no início dos anos trinta.

Gaddum e Bliss assumem a distribuição normal como uma trivialidade, conferindo maior importância à transformação logarítmica do estímulo. Os seus artigos marcam o aparecimento de um paradigma fundamental na área "bio-assay". Ambos aderem firmemente ao modelo clássico, onde o estímulo é decisivo e as respostas são aleatórias devido à variabilidade dos níveis de tolerância individual.

Bliss introduziu o termo "probit", diminuitivo de "probability unit", como uma escala conveniente para os desvios normais. No entanto, abandona rapidamente este conceito em favor de uma nova definição, segundo a qual para qualquer frequência (relativa), f, existe um equivalente desvio normal, $\tilde{Z}$, tal que a função distribuição normal em $\tilde{Z}$ é igual a f. $\tilde{Z}$ é a solução da equação

$$f = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\tilde{Z}} e^{-\frac{1}{2u^2}} du. \tag{2.5}$$

O probit da frequência f é equivalente ao desvio normal $\tilde{Z}$, ou de $\tilde{Z}$ acrescido de 5, o que assegura o facto do probit ser quase sempre positivo. Neste modelo os probits das frequências estão linearmente relacionados com o logaritmo do estímulo.

A aceitação do modelo probit foi sem dúvida facilitada pela quantidade de artigos de Bliss, que publicou regularmente até aos anos cinquenta. Também Finney e Gaddum contribuíram grandemente para essa aceitação. Após a publicação da monografia de Finney, em 1947, a análise probit foi rapidamente adoptada como um método para descrever a relação entre uma variável resposta binária e uma ou mais covariáveis.

[5]John Henry Gaddum (1900-1965) - estudou Medicina em Cambridge mas chumbou nos exames finais. Dedicou-se posteriormente à Farmacologia tendo trabalhado no "National Institute for Medical Research". Foi Professor em Farmacologia no Cairo, Londres e Edimburgo. A "British Pharmacological Society"atribui anualmente o "Gaddum Memorial Prize"na área da pesquisa farmacêutica.

[6]Charles Ittner Bliss (1899-1979) - estudou Entomologia na "Ohio State University". Trabalhou no "U.S. Department of Agriculture"até 1933, data em que este departamento foi extinto. Passou dois anos em Londres a estudar Estatística com R.A. Fisher, o qual lhe arranjou uma colocação como Estatístico em Leninegrado, onde viveu entre 1936 e 1938. Voltou para os Estados Unidos para a "Connecticut Agricultural Experiment Station", combinando o trabalho de investigador com a docência em Yale desde 1942 até à sua reforma. Foi fundamental na criação da Sociedade Biométrica.

A clássica monografia sobre a distribuição lognormal de Aitchison e Brown (1957) trouxe a análise probit ao conhecimento de muitos economistas, aumentando a sua aplicação nesta área.

Segundo J. S. Cramer a introdução da *função logística* como alternativa à função de probabilidade normal deve-se a um só homem, Joseph Berkson. Durante a década de trinta Berkson publicou inúmeros artigos na área da Medicina e Saúde Pública, mas em 1944 dedica a sua atenção à metodologia estatística em "bio-assay"(ver Berkson, 1944), propondo o uso da *função logística* em detrimento da função normal (2.5). Introduzindo o termo "*logit*"em analogia ao que fez Bliss ao introduzir o termo probit.

Berkson definiu *logit* como o inverso da *função logística* (equação 2.1)

$$\text{logit}(P) = log\frac{P}{1-P} = Z \tag{2.6}$$

a qual é, obviamente, muito mais simples que a definição de probit.

O debate sobre a questão "*logit* versus probit"é lançado por Berkson (ver Berkson, 1951), que simultaneamente ataca o método da máxima verosimilhança advogando a favor do método dos mínimos quadrados (ver Berkson, 1980). Entre 1944 e 1980 escreve inúmeros artigos sobre ambos os assuntos, adoptando um estilo provocativo e criando desta forma alguma controvérsia ao seu redor.

Foi provavelmente Wilson, em 1943, o primeiro a publicar um artigo envolvendo uma aplicação da função logística em "bio-assay"; no entanto foi Berkson quem mais lutou para o desenvolvimento e implementação da *metodologia logit* nessa área.

A proposta de Berkson não foi bem recebida pela Fundação Biométrica. Em primeiro lugar, a *metodologia logit* foi considerada como um método inferior, pouco conceituado, pois ao contrário do modelo probit não tinha qualquer "relação"com uma distribuição normal. Berkson tinha noção desta suposta "imperfeição"e tentou resolvê-la, não tendo no entanto conseguido ser muito convincente. Por esta altura ninguém, nem mesmo Berkson, parecia ter identificado o poder formidável das propriedades analíticas da *função logística*.

Em segundo lugar, o apoio de Berkson à metodologia *logit* foi prejudicado pelo seu acérrimo ataque ao método de estimação da máxima verosimilhança, elegendo preferencialmente o método dos mínimos quadrados.

Quando o debate ideológico sobre a aplicação do *modelo logit* versus modelo probit na área "bio-assay"esmoreceu, por volta de 1960, a *metodologia logit* foi largamente adoptada e a sua origem esquecida. Os primeiros desenvolvimentos tiveram lugar entre finais da década de cinquenta e princípio da década de sessenta na área da Epidemiologia.

As vantagens analíticas da transformação *logit* como um método para resolver problemas em que a variável resposta é binária rapidamente foram reconhecidas. Cox foi dos primeiros a explorar estas possibilidades, escrevendo inúmeros artigos durante a década de sessenta, culminando com um importante livro publicado em 1969.

A aplicação do *modelo logit* em "bio-assay"é generalizada através da regressão logística, onde variáveis resposta binárias são relacionadas com um variado número de covariáveis. Continuava-se, no entanto, sem encontrar uma forte fundamentação

Tabela 2.1: **No. de artigos contendo a palavra "logit"ou "probit"**

Ano	probit	logit
1935-39	6	-
1940-44	3	1
1945-49	22	6
1950-54	50	15
1955-59	53	23
1960-64	41	27
1965-69	43	41
1970-74	48	61
1975-79	45	72
1980-84	93	147
1985-89	98	215
1990-94	127	311

teórica.

Na área da Epidemiologia, os estudos de caso-controlo foram pioneiros na utilização do *modelo logit*. Mais tarde surge a ligação entre a *metodologia logit* e a Análise Discriminante, e com os modelos loglineares em geral.

A ascensão do *modelo logit* na literatura estatística é ilustrada na Tabela (2.1) (ver Cramer, 2002), a qual mostra o número de artigos que contêm a palavra "*logit*"ou "probit".

Como podemos observar, entre 1935 e 1985 o número de artigos aumentou oito vezes. Até 1970 os números mostram a predominância do modelo probit, mas a partir dessa data o *modelo logit* tomou a dianteira sem mais a deixar.

No que respeita à estimação dos modelos *logit* e probit, o método de máxima verosimilhança passou a ser norma quando rotinas computacionais são incluídas na maior parte dos "packages"estatísticos. Quando da publicação do primeiro livro de Hosmer e Lemeshow, em 1989, o uso destas rotinas era já um dado adquirido.

Concluímos, assim, que das duas causas advogadas por Berkson, a estimação por mínimos quadrados foi suplantada pela revolução computacional, enquanto que a transformação *logit* saiu sem dúvida triunfante.

Durante muito tempo a regressão logística, quer no contexto binário quer multinomial, foi usada como uma técnica, uma simples ferramenta sem fundamentação subjacente e consequentemente sem uma interpretação característica. Mas, em 1973, McFadden, que trabalhava como consultor num projecto público sobre transportes na Califórnia, estabeleceu uma ligação entre o *modelo logit* multinomial e a teoria da escolha discreta. Este trabalho forneceu uma fundamentação teórica para o *modelo logit* muito mais profunda do que qualquer teoria evidenciada no uso do modelo probit. Este trabalho valeu a McFadden o prémio Nobel em Economia no ano 2000 (ver McFadden, 2001).

Caracterização e Propriedades

Como temos vindo a referir, são inúmeras as áreas onde a variável dependente é binária. Nestes casos, os modelos que são adequados para os casos em que a variável dependente assume valores contínuos num determinado intervalo podem originar resultados difíceis de interpretar. O *modelo logit* é um dos modelos preferenciais para resolver este tipo de problema.

Comecemos por considerar o modelo de regressão linear

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.7}$$

onde

- y é uma variável binária que assume os valores 1 se o acontecimento ocorre e 0 caso contrário;
- β_0 é o coeficiente do termo constante;
- β_1 é o coeficiente da variável independente, que reflecte a influência de x em relação à probabilidade de interesse;
- x é a variável independente;
- ε é o coeficiente do erro.

O uso do modelo de regressão linear fornece, de uma forma geral, os resultados correctos em termos de sinal e do nível de significância dos coeficientes. Os problemas deste modelo, quando a variável resposta é binária, surgem ao nível das probabilidades estimadas e são os seguintes:

1. os coeficientes do erro são heterocedásticos[7], $\text{Var}(\varepsilon) = p(1-p)$, onde p é a probabilidade do acontecimento ocorrer, isto é, $p = \text{Prob}(y = 1)$. Uma vez que p depende de x, a hipótese clássica da regressão linear de que o erro não depende de x é violada;
2. o erro ε não tem distribuição normal, uma vez que p assume apenas dois valores, violando outro pressuposto clássico da regressão linear;
3. as probabilidades estimadas podem ser superiores a 1 ou inferiores a 0, o que teoricamente é inaceitável e na prática revela-se um problema sempre que necessitamos de usar esses valores.

É aqui que o *modelo de regressão logística* surge como solução de todos estes problemas. Comecemos por definir o *modelo logit* simples, isto é, o modelo constituído por uma única covariável

$$y = \text{logit}(p) = ln\frac{p}{1-p} = \beta_0 + \beta_1 x + \varepsilon \tag{2.8}$$

onde

[7] A heterocedasticidade ocorre quando a variância da variável dependente é diferente para diferentes valores das variáveis independentes.

- ln representa o logaritmo neperiano;
- $p = \text{Prob}(y = 1)$;
- $p/(1-p)$ é o rácio de probabilidade ou "odds ratio";
- $ln[p/(1-p)]$ é o logaritmo do "odds ratio"ou logit de p.

O *modelo logit* é simplesmente uma transformação não linear da clássica regressão linear. A distribuição logística é uma função distribuição em forma de "S", similar à distribuição normal, mas mais fácil de trabalhar na maioria das aplicações. A distribuição *logit* restringe as probabilidades estimadas ao intervalo $[0, 1]$.

A probabilidade estimada é dada por

$$p = \frac{e^{\widetilde{\beta}_0 + \widetilde{\beta}_1 x}}{1 + e^{\widetilde{\beta}_0 + \widetilde{\beta}_1 x}} \tag{2.9}$$

ou

$$p = \frac{1}{1 + e^{(-\widetilde{\beta}_0 - \widetilde{\beta}_1 x)}} \tag{2.10}$$

com $\widetilde{\beta}_0$ e $\widetilde{\beta}_1$ os coeficientes ajustados.

Com esta forma funcional tem-se que:

- Se $\beta_0 + \beta_1 x = 0$ então $p = 0,50$;
- Se $\beta_0 + \beta_1 x \nearrow +\infty$ então $p \nearrow 1$;
- Se $\beta_0 + \beta_1 x \searrow -\infty$ então $p \searrow 0$.

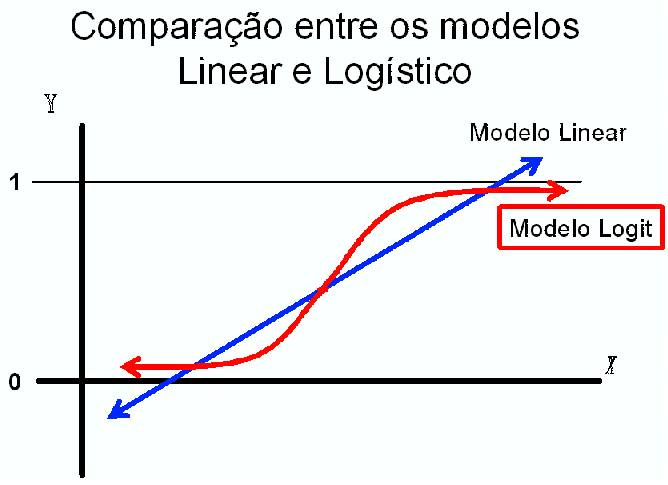

Figura 2.2: Regressão Linear vs Regressão Logística.

Como em qualquer outro modelo, também no *modelo logit* os coeficientes estimados devem ser interpretados com atenção. Por exemplo, o coeficiente angular β_1, que no modelo linear representa o impacto da variável x sobre a variável y, no *modelo logit* é interpretado como a proporção entre a variação do *logit* e a variação da variável controlada.

Para apresentarmos a teoria da estimação na forma mais geral, comecemos por definir o *modelo de regressão logística múltiplo*.

Suponhamos que temos um conjunto de m variáveis independentes, que vamos representar por

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_m)^T.$$

Sendo

$$p(\boldsymbol{x}) = \mathrm{P}(\boldsymbol{Y} = \mathbf{1}|\boldsymbol{x}),$$

o *logit* do modelo de regressão logística múltipla é dado pela equação

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m.$$

Assim,

$$p(\boldsymbol{x}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m}} = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}}$$

tendo-se o modelo

$$\boldsymbol{Y} = \mathrm{logit}(p(\boldsymbol{x})) = ln\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m. \tag{2.11}$$

Não podemos terminar esta secção sem fazer uma chamada de atenção. Na construção do modelo de regressão logística é necessário ter em consideração o tipo de variáveis independentes com que estamos a trabalhar. Se alguma das variáveis independentes for nominal, não é correcta a sua inclusão no modelo como se fosse numérica. Nesta situação, recorre-se a variáveis auxiliares, conhecidas como variáveis "dummy". Este assunto pode ser aprofundado em Hosmer e Lemeshow (2000).

Estimação dos Parâmetros

Suponhamos que temos uma amostra com n observações, $(\boldsymbol{x_i}, y_i)$ $i = 1, \ldots, n$, onde y_i representa o valor da i-ésima resposta e $\boldsymbol{x_i}$ o vector dos valores das covariáveis para a i-ésima observação. Não podemos esquecer que a variável resposta assume dois valores 0 ou 1, representando a ausência ou presença de uma dada característica. Para ajustar o *modelo logit* é necessário estimar os valores dos parâmetros desconhecidos $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)^T$.

Na regressão linear, o método mais usado na estimação é o conhecido método dos mínimos quadrados, no qual se escolhem os valores dos parâmetros que minimizam a soma dos quadrados dos desvios. Admitindo-se que o vector médio dos erros é nulo

e que a respectiva matriz de covariância é $\sigma^2 \boldsymbol{I}_n$, este método, debaixo de condições bastante gerais, produz estimadores lineares centrados com variância mínima. Infelizmente, quando aplicado a um modelo cuja variável resposta é dicotómica, os estimadores obtidos não apresentam estas propriedades.

A estimação e consequente interpretação dos parâmetros do *modelo logit* são baseadas nos seguintes princípios:

1. a independência das n observações;
2. a linearidade do *logit* relativamente às variáveis independentes;
3. a existência de duas únicas possibilidades de resposta;
4. a garantia de existirem sempre alguns "sucessos"e alguns "insucessos";
5. a dimensão da população ser suficientemente grande por forma a que, sempre que uma observação seja seleccionada, as probabilidades de "sucesso"/ "insucesso"não sejam alteradas.

Segundo Lunneborg (1994), os dois primeiros princípios são comuns ao modelo de regressão linear, enquanto que os três últimos substituem as hipóteses de normalidade e de homocedasticidade do modelo linear. No modelo de regressão linear, a resposta difere apenas no valor da média, no *modelo logit* a resposta está condicionada pela probabilidade de "sucesso".

No modelo de regressão logística os parâmetros são estimados iterativamente. Começamos com uma solução inicial, verificamos se o modelo se ajusta às observações e continuamos enquanto obtivermos uma diminuição significativa da soma dos quadrados dos resíduos.

O método da máxima verosimilhança é, sem dúvida, o método mais utilizado na estimação dos parâmetros do modelo de regressão logística, talvez por ser o mais "conhecido", talvez porque a maioria dos "packages"estatísticos o têm rotinado. Contudo, existem outros dois métodos que também têm sido utilizados na estimação destes parâmetros. São eles o método não-iterativo de mínimos quadrados ponderados e o método baseado na função discriminante (ver Hosmer e Lemeshow, 2000).

Convém referir que os estimadores de máxima verosimilhança são usualmente calculados recorrendo a algoritmos iterativos de mínimos quadrados ponderados, pelo que de alguma forma acabam por ser estimadores de "mínimos quadrados"(ver Hosmer Lemeshow, 2000).

Método da Máxima Verosimilhança

Vamos admitir que as observações têm distribuição discreta. Para adaptar a nossa discussão ao caso em que existe densidade, basta substituir a função de probabilidade pela densidade.

De uma forma bem geral, podemos dizer que o método de máxima verosimilhança produz valores para os parâmetros desconhecidos que maximizam a probabilidade de obter o conjunto de dados observados.

O primeiro passo para a determinação dos estimadores de máxima verosimilhança passa pela construção da função de verosimilhança, que representa a probabilidade associada à informação obtida como função dos parâmetros desconhecidos. Os estimadores de máxima verosimilhança destes parâmetros são os valores que maximizam esta função, para que no fim sejam os que mais se aproximam dos dados observados.

A forma mais conveniente de expressar a contribuição de cada par $(\boldsymbol{x_i}, y_i)$, $i = 1, \ldots, n$, é dada pela equação

$$p(\boldsymbol{x_i}|\boldsymbol{\beta})^{y_i}[1 - p(\boldsymbol{x_i}|\boldsymbol{\beta})]^{1-y_i}. \tag{2.12}$$

Como, por hipótese, as observações são independentes, a função de verosimilhança é obtida por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(\boldsymbol{x_i}|\boldsymbol{\beta})^{y_i}[1 - p(\boldsymbol{x_i}|\boldsymbol{\beta})]^{1-y_i}. \tag{2.13}$$

Há, agora, que determinar os valores que maximizam esta função, o que intuitivamente leva a escolher os valores, em função das observações, que tornam mais provável a amostra observada. Devemos então derivar a função verosimilhança em ordem aos parâmetros do modelo e igualar a zero. Na prática, porque facilita os cálculos e porque são questões equivalentes, aplica-se a função logaritmo à função verosimilhança obtendo-se:

$$\ell(\boldsymbol{\beta}) = ln(L(\boldsymbol{\beta})) = \sum_{i=1}^{n}(y_i ln[p(\boldsymbol{x_i}|\boldsymbol{\beta})] + (1 - y_i) ln[1 - p(\boldsymbol{x_i}|\boldsymbol{\beta})]) =$$

$$= \sum_{i=1}^{n} y_i ln \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}}}{1 - e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}}} + (1 - y_i) ln[1 - \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}}}{1 - e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}}}] =$$

$$= \sum_{i=1}^{n} y_i\beta_0 + y_i\beta_1 x_{i1} + \cdots + y_i\beta_m x_{im} - ln(1 + e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}}). \tag{2.14}$$

As equações de verosimilhança são

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta_0}} = \sum_{i=1}^{n}[y_i - p(\boldsymbol{x_i}|\boldsymbol{\beta})]$$

e

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}[y_i - p(\boldsymbol{x_i}|\boldsymbol{\beta})]$$

para $j = 1, 2, \ldots, m$.

Como referimos anteriormente, a solução destas equações requer a utilização de um método iterativo. Um dos mais utilizados é o método de Newton ou Newton-Raphson.

A matriz de covariância dos estimadores de máxima verosimilhança é dada, ver Amemiya (1985) ou Davidson and MacKinnon (1993), pelo inverso da matriz de informação

$$\mathbf{I}(\boldsymbol{\beta}) = -[\mathrm{E}(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_{j+1}\partial\beta_{l+1}})]$$

com $j, l = 0, 1, \ldots, m$.

Neste caso, as derivadas de segunda ordem são

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j^2} = -\sum_{i=1}^{n} x_{ij+1}^2 p_i(1-p_i)$$

e

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_l} = -\sum_{i=1}^{n} x_{ij+1}x_{il+1} p_i(1-p_i)$$

onde $j, l = 0, 1, \ldots, m$, $j \neq l$ e $p_i = p(x_i)$ são constantes, pelo que são iguais nos respectivos valores médios.

Representamos as variâncias e as covariâncias dos estimadores de máxima verosimilhança por $\mathrm{Var}(\widehat{\beta}_j)$, $j = 0, 1, \ldots, m$, e por $\mathrm{Cov}(\widehat{\beta}_j, \widehat{\beta}_l)$, $j, l = 0, 1, \ldots, m$ e $j \neq l$, respectivamante. Como referimos, essas variâncias e covariâncias são os elementos da matriz $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})$.

Por sua vez, as $\widehat{\mathrm{Var}}(\widehat{\beta}_j)$ e as $\widehat{\mathrm{Cov}}(\widehat{\beta}_j, \widehat{\beta}_l)$, $j, l = 0, 1, \ldots, m$ e $j \neq l$, serão os estimadores dessas variâncias e covariâncias.

Usando notação matricial, podemos escrever

$$\widehat{\boldsymbol{I}}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}$$

e

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = [\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}]^{-1}$$

onde $\boldsymbol{X}$ e $\boldsymbol{V}$ são as matrizes

$$\boldsymbol{X}_{n\times(m+1)} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

e

$$\boldsymbol{V}_{n\times n} = \begin{pmatrix} \widehat{p_1}(1-\widehat{p_1}) & 0 & \cdots & 0 \\ 0 & \widehat{p_2}(1-\widehat{p_2}) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \widehat{p_n}(1-\widehat{p_n}) \end{pmatrix}$$

onde $\widehat{p_i} = \widehat{p}(x_i)$, $i = 1, 2, \ldots, n$, representam os valores ajustados do modelo logit.

O cálculo desta matriz é fundamental para podermos posteriormente testar a significância dos parâmetros estimados.

Testes para a Significância dos Parâmetros do Modelo

Após estimar os parâmetros, deve-se testar a significância das variáveis que compõem esse modelo.

Esta tarefa envolve a formulação de testes de hipóteses que permitam determinar quais das variáveis independentes estão significativamente relacionadas com a variável resposta.

Vamos apresentar, de forma sucinta, os testes mais utilizados na regressão logística.

Segundo Hosmer e Lemeshow (2000), a construção de um teste à significância dos parâmetros associados às variáveis deve ter em conta a seguinte questão:

"*Does the model that includes the variables tell us more about the outcome variable than the model that does not include the variables?*"

A resposta a esta questão obtém-se comparando os valores observados para a variável resposta com os valores estimados através dos modelos, com e sem algumas das variáveis.

Na regressão logística, a comparação entre valores estimados e valores observados é baseada na função log-verosimilhança, através da estatística

$$D = -2ln[\frac{L_a}{L_s}]$$

que é designada por "Deviance", onde L_a representa o valor da função de verosimilhança do modelo actual e L_s o valor da função de verosimilhança do modelo saturado[8].

O teste baseado nesta estatística é denominado por **Teste de razão de verosimilhanças**.

Tendo em conta a equação (2.14), a estatística do teste fica com o aspecto

$$D = -2\sum_{i=1}^{n}[y_i ln(\frac{\widehat{p_i}}{y_i}) + (1-y_i)ln(\frac{1-\widehat{p_i}}{1-y_i})].$$

Quando os valores observados se afastarem muito dos valores estimados, o valor desta estatística será grande.

Para testar a hipótese H_0 de que s dos m, com $s < m$, parâmetros do modelo são iguais a zero, isto é, que s covariáveis não são significativas, contra a hipótese alternativa de existir pelo menos uma que o é, construímos uma nova estatística baseada na diferença do valor da "Deviance", D, com e sem as s variáveis. Esta nova estatística é definida por

$$G = D(\text{modelo sem as } s \text{ variáveis}) - D(\text{modelo com as } s \text{ variáveis})$$

ou seja

$$G = -2ln[\frac{\text{valor da verosimilhança sem as } s \text{ variáveis}}{\text{valor da verosimilhança com as } s \text{ variáveis}}].$$

[8]Um modelo diz-se saturado quando contém tantos parâmetros quantas as observações.

Sob a hipótese H_0, esta estatística distribui-se como um qui-quadrado central com s graus de liberdade[9].

O valor da estatística G serve para fazer comparações entre os diversos modelos que vamos obtendo ao retirar uma ou mais variáveis, permitindo alguma elasticidade por forma a conseguir obter o modelo que fornece um menor valor do $p-value$. No entanto, na prática, este processo revela-se na grande maioria das vezes algo moroso.

O teste de razão de verosimilhanças é, segundo Hosmer e Lemeshow (2000), o teste mais utilizado na regressão logística, opinião que é partilhada por muitos outros autores.

Para terminar esta secção, vamos apenas referenciar dois outros testes também utilizados no modelo logístico.

O **teste de Wald** é baseado na estatística

$$W = \widehat{\boldsymbol{\beta}}^T [\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})]^{-1} \widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^T (\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X}) \widehat{\boldsymbol{\beta}}$$

que tem distribuição de qui-quadrado com $(m+1)$ graus de liberdade, sob a hipótese nula de que os $(m+1)$ parâmetros são iguais a zero.

Testes para os m parâmetros enviesados são obtidos eliminando o parâmetro $\widehat{\beta}_0$ do vector $\widehat{\boldsymbol{\beta}}$ e a correspondente linha (primeira ou última) e coluna (primeira ou última) da matriz $(\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})$.

Como este teste requer a realização de operações entre matrizes e de obter o vector $\widehat{\boldsymbol{\beta}}$, não parece haver vantagens deste método relativamente ao teste de razão de verosimilhanças.

O **teste "Score"** para a significância do modelo é baseado na distribuição das m derivadas da função log-verosimilhança, $\ell(\boldsymbol{\beta})$, em ordem ao vector $\boldsymbol{\beta}$.

Em termos de cálculos, este teste está equiparado ao teste de Wald.

A temática abordada nesta secção pode ser aprofundada em Cox e Hinkley (1974) ou Dobson (1990).

Intervalos de Confiança

Consideremos a seguinte expressão genérica para o estimador do *modelo logit* com m covariáveis

[9]Esta afirmação tem por base o Teorema de Wilks, segundo o qual, sob as condições de regularidade (ver Rohagti, 1976), $(-2ln\Lambda(x))$ tem distribuição assimptótica qui-quadrado com número de graus de liberdade correspondentes à diferença entre o número de parâmetros independentes em Θ e em Θ_0 onde

$$\Lambda(x) = \frac{\sup\limits_{\theta \in \Theta_0} f_\theta(x_1, \ldots, x_m)}{\sup\limits_{\theta \in \Theta} f_\theta(x_1, \ldots, x_m)}$$

sendo

$$f_\theta = \prod_{i=1}^{n} f(x_i, \theta)$$

quando se testa a hipótese $H_0 : \theta \in \Theta_0$ contra a hipótese alternativa $H_1 : \theta \in \Theta_1$, onde Θ_0 e Θ_1 constituem uma partição de Θ.

$$\widehat{g}(\boldsymbol{x}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_m x_m$$

ou, usando notação vectorial,

$$\widehat{g}(\boldsymbol{x}) = \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}$$

onde $\widehat{\boldsymbol{\beta}}^T = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_m)$ denota o estimador dos $m+1$ parâmetros e o vector $\boldsymbol{x}^T = (x_0, x_1, x_2, \ldots, x_m)$ tem como componentes a constante $x_0 = 1$ e as m covariáveis que compõem o modelo.

O estimador da variância do estimador do *logit* é

$$\widehat{\mathrm{Var}}[\widehat{g(\boldsymbol{x})}] = \sum_{j=0}^{m} x_j^2 \widehat{\mathrm{Var}}(\widehat{\beta}_j) + \sum_{j=0}^{m} \sum_{k=j+1}^{m} 2 x_j x_k \widehat{\mathrm{Cov}}(\widehat{\beta}_j, \widehat{\beta}_k)$$

ou, usando notação matricial,

$$\widehat{\mathrm{Var}}[\widehat{g(\boldsymbol{x})}] = \boldsymbol{x}^T \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) \boldsymbol{x} = \boldsymbol{x}^T [\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X}]^{-1} \boldsymbol{x}.$$

Na prática, todos estes cálculos são efectuados recorrendo a software estatístico, o que elimina todo o peso associado ao cálculo matricial.

2.1.2 Método de Mínimos Quadrados na Regressão Não Linear

A regressão não linear pelo método dos mínimos quadrados é uma extensão do método linear que permite a aplicação desta metodologia a uma classe mais geral de funções.

Ao contrário do que acontece no modelo linear, no modelo não linear são poucas as limitações que existem na forma como os parâmetros podem ser expressos no modelo. No entanto, o processo de estimação destes parâmetros é, em termos conceptuais, muito semelhante à estimação no modelo linear.

Como o próprio nome indica, um modelo não linear é qualquer modelo cuja forma genérica é dada por

$$\boldsymbol{Y}_i = f(\boldsymbol{X}_i, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_i \tag{2.15}$$

com

$$\boldsymbol{X}_i = \begin{pmatrix} \mathrm{X}_{i1} \\ \mathrm{X}_{i2} \\ \vdots \\ \mathrm{X}_{im-1} \end{pmatrix} \quad \text{e} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{m-1} \end{pmatrix}$$

onde $f(\boldsymbol{X}_i, \boldsymbol{\beta})$ representa o valor esperado para o i-ésimo caso, função não linear dos parâmetros desconhecidos.

A grande vantagem do método dos mínimos quadrados, relativamente a outros, é conduzir, em muitos caos, a bons estimadores. Saliente-se ainda o grande desenvolvimento da respectiva teoria.

A principal dificuldade na aplicação deste método é a necessidade de recorrer a técnicas iterativas. Estas técnicas são sensíveis à escolha dos valores iniciais.

Em particular, uma escolha errada dos valores iniciais pode conduzir a um mínimo local e não a um mínimo global, como se pretende na estimação pelo método dos mínimos quadrados.

Consideremos novamente a expressão genérica do modelo não linear

$$\boldsymbol{Y}_i = f(\boldsymbol{X}_i, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_i.$$

Tal como no modelo de regressão linear, procura-se minimizar

$$\boldsymbol{Q} = \sum_{i=1}^{n} (\boldsymbol{Y}_i - f(\boldsymbol{X}_i, \boldsymbol{\beta}))^2.$$

em ordem a $\boldsymbol{\beta}$.

As derivadas parciais de $\boldsymbol{Q}$ relativamente aos parâmetros β_j, $j = 0, 1, \ldots, m-1$, são dadas por

$$\frac{\partial \boldsymbol{Q}}{\partial \beta_j} = \sum_{i=1}^{n} -2(\boldsymbol{Y}_i - f(\boldsymbol{X}_i, \boldsymbol{\beta}))[\frac{\partial f(\boldsymbol{X}_i, \boldsymbol{\beta})}{\partial \beta_j}],$$

o que nos leva a resolver o sistema

$$\sum_{i=1}^{n} \boldsymbol{Y}_i [\frac{\partial f(\boldsymbol{X}_i, \boldsymbol{\beta})}{\partial \beta_j}]_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = -\sum_{i=1}^{n} f(\boldsymbol{X}_i, \boldsymbol{\beta})[\frac{\partial f(\boldsymbol{X}_i, \boldsymbol{\beta})}{\partial \beta_j}]_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = 0. \tag{2.16}$$

A solução do mesmo será

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_{m-1} \end{pmatrix}. \tag{2.17}$$

Como (2.16) é um sistema de equações não lineares, recorremos a métodos iterativos para o resolver. Destes métodos, o de Gauss-Newton pode muitas vezes ser utilizado com vantagem.

Método de Gauss-Newton

O método de Gauss-Newton é um exemplo dum procedimento de procura numérica directa. Este método assenta em expansões de Taylor

$$f(\boldsymbol{X}_i, \boldsymbol{\beta}) \approx f(\boldsymbol{X}_i, \boldsymbol{b}^{(0)}) + \sum_{j=0}^{m-1} [\frac{\partial f(\boldsymbol{X}_i, \boldsymbol{\beta})}{\partial \beta_j}]_{\boldsymbol{\beta}=\boldsymbol{b}^{(0)}} (\beta_j - b_j^{(0)}) \tag{2.18}$$

para obter uma aproximação linear do modelo.

Aplica-se, então, o método dos mínimos quadrados à aproximação linear.

Partindo dos valores iniciais

$$b_0^{(0)}, b_1^{(0)}, \ldots, b_{m-1}^{(0)}$$

realiza-se uma primeira optimização, obtendo-se os valores

$$b_0^{(1)}, b_1^{(1)}, \ldots, b_{m-1}^{(1)}$$

que são tomados como ponto de partida para a iteração seguinte.

No fim de cada iteração, obtém-se a soma dos quadrados dos resíduos. Esta começa por diminuir, passando em seguida a oscilar devido a erros de aproximação.

Particularizando, consideremos

$$f_i^{(0)} = f(\boldsymbol{X}_i, \boldsymbol{b}^{(0)}),$$

$$\delta_j^{(0)} = (\beta_j - b_j^{(0)})$$

$$D_{ij}^{(0)} = [\frac{\partial f(\boldsymbol{X}_i, \boldsymbol{\beta})}{\partial \beta_j}]_{\boldsymbol{\beta}=\boldsymbol{b}^{(0)}}$$

o que permite reescrever a aproximação (2.18) como

$$f(\boldsymbol{X}_i, \boldsymbol{\beta}) \approx f_i^{(0)} + \sum_{j=0}^{m-1} D_{ij}^{(0)} \delta_j^{(0)}. \tag{2.19}$$

Então, uma aproximação do modelo $\boldsymbol{Y}_i = f(\boldsymbol{X}_i, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_i$ é dada por

$$\boldsymbol{Y}_i \approx f_i^{(0)} + \sum_{j=0}^{m-1} D_{ij}^{(0)} \delta_j^{(0)} + \varepsilon_i, \ i = 1, \ldots, n. \tag{2.20}$$

Passando $f_i^{(0)}$ para o lado esquerdo da equação e representando a diferença $Y_i - f_i^{(0)}$ por $Y_i^{(0)}$, obtemos

$$Y_i^{(0)} \approx \sum_{j=0}^{m-1} D_{ij}^{(0)} \delta_j^{(0)} + \varepsilon_i, \ i = 1, \ldots, n. \tag{2.21}$$

Cada coeficiente de regressão $\delta_j^{(0)}$ representa a diferença entre os verdadeiros parâmetros da regressão e as estimativas iniciais dos mesmos. Assim, os $\delta_j^{(0)}$ representam uma correcção que deve ser feita nos coeficientes de regressão iniciais. O objectivo de ajustar o modelo de regressão linear (2.21) é estimar os $\delta_j^{(0)}$ e usar essas estimativas para corrigir as estimativas iniciais dos parâmetros.

Escrevamos o modelo (2.21) na forma matricial

$$\boldsymbol{Y}^{(0)} \approx \boldsymbol{D}^{(0)} \boldsymbol{\delta}_j^{(0)} + \boldsymbol{\varepsilon} \tag{2.22}$$

onde

$$\boldsymbol{Y}^{(0)}_{n\times 1} = \begin{pmatrix} Y_1 - f_1^{(0)} \\ \vdots \\ Y_n - f_n^{(0)} \end{pmatrix},$$

$$\boldsymbol{D}^{(0)}_{n\times m} = \begin{pmatrix} D_{10}^{(0)} & \cdots & D_{1m-1}^{(0)} \\ \vdots & \cdots & \vdots \\ D_{n0}^{(0)} & \cdots & D_{nm-1}^{(0)} \end{pmatrix},$$

$$\boldsymbol{\delta}^{(0)}_{m\times 1} = \begin{pmatrix} \delta_0^{(0)} \\ \vdots \\ \delta_{m-1}^{(0)} \end{pmatrix}$$

e

$$\boldsymbol{\varepsilon}_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Podemos agora estimar os parâmetros $\boldsymbol{\delta}^{(0)}$ pelo método de mínimos quadrados

$$\boldsymbol{d}^{(0)} = (\boldsymbol{D}^{(0)T}\boldsymbol{D}^{(0)})^{-1}\boldsymbol{D}^{(0)T}\boldsymbol{Y}^{(0)}$$

e em seguida determinar os coeficientes de regressão estimados corrigidos

$$\boldsymbol{b}^{(1)} = \boldsymbol{b}^{(0)} + \boldsymbol{d}^{(0)}.$$

Devemos, agora, verificar se os coeficientes de regressão corrigidos representam uma melhoria na direcção desejada.

No início tínhamos a soma de quadrados de resíduos

$$SQE^{(0)} = \sum_{i=1}^{n}(Y_i - f(X_i, b^{(0)}))^2 = \sum_{i=1}^{n}(Y_i - f_i^{(0)})^2$$

e no final da primeira iteração a soma será

$$SQE^{(1)} = \sum_{i=1}^{n}(Y_i - f(X_i, b^{(1)}))^2 = \sum_{i=1}^{n}(Y_i - f_i^{(1)})^2.$$

Se o algoritmo de Gauss-Newton for a escolha correcta para o caso em estudo, $SQE^{(1)}$ será menor do que $SQE^{(0)}$.

Repete-se o procedimento enquanto as somas de quadrados dos resíduos forem diminuindo.

Para um estudo mais aprofundado deste ou de outro método numérico sugerimos a seguinte bibliografia: Kelley (2003), Faires *et al.* (2002) e Pina (1995).

2.1.3 Meta-Teorema de Fisher

Para melhor nos apercebermos da importância deste teorema de Fisher, nada melhor que começarmos por apresentar o Teorema do Limite Central.

Teorema do Limite Central é, na realidade, o nome comum para uma série de teoremas limite na teoria das probabilidades segundo os quais, sob certas condições[10], a soma (ou outras funções) de um grande número de variáveis aleatórias estandardizadas independentes (ou fracamente dependentes) tem distribuição aproximadamente Normal.

Teorema do Limite Central (versão clássica)

Consideremos a sequência $X_1, \ldots, X_n, \ldots$ de variáveis aleatórias com valor médio finito $\mathrm{E}[X_i] = \mu_i$, variância finita $\mathrm{Var}[X_i] = \sigma_i^2$, e $S_n = \sum_{i=1}^{n} X_i$.

Considerando

$$\mathrm{A}_n = \mathrm{E}[S_n] = \mu_1 + \cdots + \mu_n \quad \text{e} \quad \mathrm{B}_n = \mathrm{Var}[S_n] = \sigma_1^2 + \cdots + \sigma_n^2$$

a distribuição F_n da variável estandardizada $Z_n = \frac{S_n - A_n}{\sqrt{B_n}}$ converge para a distribuição normal estandardizada

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2} dz$$

isto é $F_n(x) \to \Phi(x)$ ou, de forma equivalente, para qualquer intervalo $]\alpha, \beta[$

$$P\{\alpha < Z_n < \beta\} = P\{A_n + \alpha\sqrt{B_n} < S_n < A_n + \beta\sqrt{B_n}\} \to \Phi(\beta) - \Phi(\alpha).$$

[10] O Teorema de Lyapunov (na teoria das probabilidades) estabelece condições suficientes para a convergência da distribuição da soma de variáveis aleatórias independentes para a distribuição normal. Mais exactamente, este teorema afirma o seguinte:

Suponhamos que as variáveis aleatórias independentes $X_1, \ldots, X_n, \ldots$, têm valores médios finitos $\mathrm{E}[X_k] = \mu_k$, variâncias finitas $\mathrm{Var}[X_k] = \sigma_k^2$, momentos absolutos $\mathrm{E}|X_k - \mu_k|^{2+\delta}$, com $\delta > 0$, e que $B_n = \sum_{k=1}^{n} \sigma_k^2$ é a variância da soma das variáveis aleatórias. Então, se para algum $\delta > 0$

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} \mathrm{E}|X_k - \mu_k|^{2+\delta}}{B_n^{1+\delta/2}} = 0,$$

a probabilidade da desigualdade

$$x_1 < \frac{\sum_{k=1}^{n} (X_k - \mu_k)}{\sqrt{B_n}} < x_2$$

tende, quando $n \to \infty$, para o limite $\frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-x^2/2} dx$ uniformemente para todos os valores de x_1 e x_2.

Este teorema foi apresentado e demonstrado por A.M. Lyapunov em 1901.

Mais tarde S.N. Bernstein, J. Lindeberg e W. Feller mostraram que estas condições são não só suficientes mas também necessárias (ver Feller, 1968 e Petrov, 1975).

Como facilmente podemos constatar através da consulta da extensa bibliografia de R.A. Fisher[11], muitas das suas ideias chave foram inspiradas em problemas Biológicos.

Para Fisher a genética Mendeliana explicava potencialmente a grande variação existente nas observações biométricas (ver Howie, 2002). Fisher mostrou um completo desinteresse pela informação individual e mostrou que a inferência em Biologia deve ser feita com base na distribuição dos valores das características Mendelianas[12] em populações onde se verificam intercruzamentos generalizados. O que vai de acordo com o seu ponto de vista de que a inferência estatística deve ser baseada na comparação das estatísticas observadas com a distribuição teórica amostral infinita.

Em 1918, Fisher publicou o artigo intitulado "*The correlation between relatives on the supposition of Mendelian inheritance*", que foi, e ainda é, considerado um marco milenar tanto na Estatística como na Biologia. Neste artigo Fisher introduz o protótipo da "Análise da Variância", bem como a síntese do Mendelismo, biometria e evolução (ver Moran and Smith, 1966).

Aldrich (1995) teceu grandes elogios a este artigo, considerando-o "*the most ambitious piece of scientific inference*".

Uma das conjecturas cruciais que Fisher apresentou neste artigo foi que as características contínuas são determinadas por um grande número de factores Mendelianos. Esta conjectura, aparentemente inócua, permitiu a Fisher argumentar que a distribuição associada às características deve ser a distribuição Normal.

Meta-Teorema de Fisher

O número de factores é virtualmente infinito, mas cada um deles tem um efeito muito pequeno e actua independentemente dos outros. A distribuição Normal assimptótica é então uma consequência do teorema do limite central para distribuições.

Uma vez assegurada a normalidade das distribuições, Fisher calculou as diversas correlações entre parentes e verificou que se encontravam, aproximadamente, de

[11] Sir Ronald Aylmer Fisher (1890-1962) foi um biólogo, geneticista e estatístico de extraordinário talento. Richard Dawkins descreveu-o como "The greatest of Darwin's successors"e o historiador Anders Hald escreveu "Fisher was a genius who almost single-handedly created the foundations for modern statistical science". Fisher inventou as técnicas da máxima verosimilhança e da análise de variância, foi pioneiro no planeamento de experiências e definiu os conceitos de suficiência, ancilaridade e informação de Fisher. Tudo isto fez com que Fisher fosse considerado a maior figura da Estatística do século XX.

[12] Características Mendelianas são características que são transmitidas por um só par de genes, isto é, que seguem sem restrições as leis de Mendel (Gregor Mendel, botânico austríaco que viveu entre 1822 e 1884), que sumariamentente afirmam que:

- Cada característica física corresponde a um único gene;
- Os genes encontram-se em pares;
- Apenas um gene de cada par é transmitido à geração seguinte por cada progenitor;
- A transmissão dos dois genes de cada par é igualmente provável;
- Algumas características são dominantes e outras são recessivas.

(Ver Pierce (2005) e Stern(1966)).

acordo com as medições efectuadas pelos biometristas.

Para um mais completo estudo deste assunto o ideal será ler Fisher (1918, 1928 e 1954).

2.2 Expansões de Edgeworth

A distribuição Normal desempenha um papel fundamental nas mais diversas áreas de aplicação da Estatística. Em muitas aplicações, por forma a extrair informações e conclusões úteis, pode ser mais importante medir os desvios da função densidade da Normal do que provar que a distribuição associada ao fenómeno em estudo está próxima da distribuição Normal.

No entanto por vezes é preciso ir além das distribuições normais utilizando-se expansões assimptóticas. Estas têm desempenhado um papel importante na Teoria Estatística:

- fornecem respostas aproximadas, relativamente simples, a problemas com distribuições exactas de difícil implementação;
- permitem a implementação de expansões de ordem superior a partir das quais simples expansões podem ser avaliadas e, sempre que necessário, melhoradas;
- têm conduzido ao desenvolvimento sistemático da teoria da aproximação à inferência estatística, com aplicações a famílias genéricas de modelos estatísticos que necessitam de formulação assimptótica.

Existem vários tipos de expansões, nomeadamente de Gram-Charlier, de Gauss-Hermite e de Edgeworth. São estas últimas que nos interessam e sobre as quais nos vamos debruçar.

Na grande maioria dos casos são as expansões de Edgeworth[13] que permitem obter melhores resultados, pois estão directamente ligadas aos momentos e cumulantes (semi-invariantes) da função densidade de probabilidade e também porque, sendo verdadeiras expansões assimptóticas, o erro da aproximação está controlado.

Uma expansão de Edgeworth representa uma função distribuição definida através de uma série constituída pelas derivadas da função densidade normal e pelos cumulantes[14]. A representação desta série é desenvolvida com base nos termos da função característica.

Sem querer aprofundar este assunto, decidimos apresentar, de uma forma geral, como obter a expressão de uma expansão de Edgeworth.

[13]Francis Ysidro Edgeworth (1845-1926) nasceu em Edgeworthstown na Irlanda e formou-se pela Universidade de Oxford em 1869.

Edgeworth obteve resultados extremamente valiosos para o desenvolvimento da Inferência Estatística, muitos dos quais só foram completamente compreendidos mais tarde. Os seus primeiros trabalhos foram sobre a aplicação da Matemática às Ciências Sociais. No seu livro "Mathematical Psychics"ele trabalhou os aspectos práticos da ética utilitarista em forma matemática.

[14]Recorde-se que o cumulante de ordem j, κ_j, é um polinómio de grau j cujos termos são momentos.

Dado, ver Bateman e Erdély (1954), a função característica da densidade normal reduzida, $z(x)$, ser $e^{-t^2/2}$, usando a fórmula da inversão obtém-se

$$z(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx}e^{-t^2/2}dt$$

logo

$$z^{<r>}(x) = \frac{d^r z}{dx^r} = \frac{1}{2\pi}\int_{-\infty}^{+\infty}(-it)^r e^{-itx}e^{-t^2/2}dt.$$

Caso a função característica, $\varphi(t)$, tenha um desenvolvimento em série de funções

$$\varphi(t) = (1 + \sum_{r=0}^{\infty} a_r(-it)^r)e^{-t^2/2}$$

a densidade correspondente será dada por

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx}(1 + \sum_{r=0}^{\infty} a_r(-it)^r)e^{-t^2/2}dt = z(x) + \sum_{r=0}^{\infty} a_r z^{<r>}(x).$$

Sendo $f(x)$ a densidade da variável aleatória X, a variável aleatória $Y = \frac{X}{\sigma}$ terá a densidade

$$q(x) = \sigma f(\sigma x) = \sigma(z(x) + \sum_{r=0}^{\infty} a_r z^{<r>}(\sigma x)),$$

o que é útil quando a variável aleatória X tem valor médio nulo e se quer trabalhar com uma variável estandardizada.

Uma variante bem conhecida desta técnica é dada pelas expansões de Edgeworth em que, ver Blimmikov e Moessner (1998), os coeficientes a_r são polinómios formados por simples combinações dos cumulantes e de polinómios de Hermite[15].

Para um estudo mais profundo desta temática permitimo-nos sugerir, Cramér (1957), Barndorff-Nielsen e Cox (1989), Blimmikov e Moessner (1998), Abramowitz e Stegum (1972), Juszkiewicz *et al.* (1995), Bernardeau e Kofman (1995) e, por último, Petrov (1962, 1975, 1995).

[15]Considerando sucessivas derivadas da função distribuição normal formamos uma série de polinómios, $p_i(x)$, que são ortogonais entre si relativamente à densidade normal, $z(x)$, isto é, tal que, para $i \neq j$

$$\int_{-\infty}^{+\infty} p_i(x)p_j(x)z(x)dx = 0.$$

Temos também que

$$\begin{cases} \frac{dz}{dx} = -xz(x) \\ \frac{d^2z}{dx^2} = (x^2-1)z(x) \\ \frac{d^3z}{dx^3} = (3x - x^3)z(x) \\ \vdots \end{cases}$$

pelo que os polinómios de Hermite, ver Kendall (1952), são

2.3 Resultados Algébricos

Nesta secção apresentamos dois resultados algébricos fundamentais para o desenvolvimento do trabalho futuro.

A matriz Inversa Generalizada

Para qualquer matriz complexa, mesmo que esta não seja quadrada, pode-se definir inversa generalizada ou pseudoinversa, sendo possível encontrar várias pseudoinversas para a mesma matriz.

Seja $\boldsymbol{A}$ uma matriz complexa de tipo $m \times n$. Uma matriz inversa generalizada de $\boldsymbol{A}$ é uma matriz $\boldsymbol{A}^{-}$, do tipo $m \times n$, tal que

$$\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{A} = \boldsymbol{A}.$$

Uma das inversas generalizadas mais utilizadas é a conhecida inversa de Moore-Penrose, que apresentamos de seguida.

A Matriz Inversa Generalizada de Moore-Penrose

Seja $\boldsymbol{B}$ uma matriz do tipo $m \times n$. Existe uma e uma só matriz $\boldsymbol{B}^{\dagger}$, do tipo $n \times m$, que satisfaz as seguintes condições:

- $\boldsymbol{B}\boldsymbol{B}^{\dagger}\boldsymbol{B} = \boldsymbol{B}$;
- $\boldsymbol{B}^{\dagger}\boldsymbol{B}\boldsymbol{B}^{\dagger} = \boldsymbol{B}^{\dagger}$;
- $(\boldsymbol{B}\boldsymbol{B}^{\dagger})^{T} = \boldsymbol{B}\boldsymbol{B}^{\dagger}$;
- $(\boldsymbol{B}^{\dagger}B)^{T} = \boldsymbol{B}^{\dagger}\boldsymbol{B}$.

A matriz $\boldsymbol{B}^{\dagger}$ é a matriz inversa generalizada de Moore-Penrose da matriz $\boldsymbol{B}$.

Esta matriz, que foi definida independentemente por Moore[16] em 1920 e Penrose[17] em 1955, tem tido inúmeras aplicações na área da Estatística. Por exemplo,

$$\begin{cases} H_0 = 1 \\ H_1 = x \\ H_2 = x^2 - 1 \\ H_3 = x^3 - 3x \\ \vdots \end{cases}$$

[16]E.H. Moore (1862-1932) introduziu a temática da inversa generalizada no período de 1910 a 1920, apresentando o problema da seguinte forma: "The effectiveness of the reciprocal of a nonsingular finite matrix in the study of properties of such matrices, makes it desirable to define if possible an analogous matrix to be associated with each finite matrix $\boldsymbol{B}$ even $\boldsymbol{B}$ is not square or, if square, is not necessarily nonsingular".

[17]Sir Roger Penrose redescobriu a inversa generalizada em 1955, provando a sua existência e estabelecendo a sua unicidade e principais propriedades. Não poderia ter sido de outra forma, pois o trabalho de Moore foi caindo no esquecimento, mesmo enquanto este foi vivo.

$$\boldsymbol{z} = \boldsymbol{B}^{\dagger}\boldsymbol{c}$$

é o caminho mais simples para obter a solução que minimiza o quadrado da soma de

$$\boldsymbol{c} - \boldsymbol{B}\boldsymbol{z}.$$

Se existir matriz inversa da matriz $(\boldsymbol{B}^T\boldsymbol{B})$, tem-se

$$\boldsymbol{B}^{\dagger} = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T$$

onde $\boldsymbol{B}^T$ é a matriz transposta de $\boldsymbol{B}$. Multiplicando ambos os membros da equação $\boldsymbol{B}\boldsymbol{z} = \boldsymbol{c}$ por $\boldsymbol{B}^T$ obtemos

$$\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{z} = \boldsymbol{B}^T\boldsymbol{c}$$

donde

$$\boldsymbol{z} = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{c} \equiv \boldsymbol{B}^{\dagger}\boldsymbol{c}.$$

Processo de Ortogonalização de Gram-Schmidt

Dado um espaço Euclidiano[18] E de dimensão finita, é sempre possível obter uma base ortonormada[19] de E. O processo de ortogonalização de Gram-Schmidt, que seguidamente descrevemos, permite obter uma base ortonormada de E a partir de uma qualquer base desse espaço.

Prova-se o seguinte resultado, cuja demonstração pode ser consultada em Luz *et al.* (2002):

Teorema 2.3.1. A partir de qualquer conjunto de vectores $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ linearmente independentes obtém-se, pelo processo de ortogonalização de Gram-Schmidt, um conjunto de vectores $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$, ortonormados.

Como consequência deste teorema, conclui-se que:

Corolário 2.3.2. Qualquer espaço Euclidiano de dimensão finita tem uma base ortonormada.

Para esta secção sugerimos Strang (1988), Bretscher (2004), Bapat (200) e Campbell e Meyer (1979).

Sir Roger Penrose tem feito contribuições fundamentais em áreas como a Física, Matemática, Geometria, Inteligência Artificial, etc. Em 1988 ganhou o prémio Wolf em Física conjuntamente com Stephen W. Hawking.

[18] Um espaço Euclidiano é um espaço vectorial onde está definido um produto interno.

[19] Uma base diz-se ortonormada se os seus vectores são ortogonais dois a dois e têm todos norma igual a um.

Tabela 2.2: **Processo de Ortogonalização de Gram-Schmidt**

Processo de Ortogonalização de Gram-Schmidt
Entrada: Base $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ do espaço Euclidiano E;
Saída: Base ortonormada $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ de E;
1º Passo: Calcular $\boldsymbol{q}_1 = \frac{\boldsymbol{x}_1}{\|\boldsymbol{x}_1\|}$ e fazer $j = 1$;
2º Passo: Se $j = n$, terminar ($\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ é uma base ortonormada de E), de contrário fazer $j = j + 1$ e ir para o 3º passo;
3º Passo: Calcular $\boldsymbol{q}'_j = \boldsymbol{x}_j - (\boldsymbol{x}_j \vert \boldsymbol{q}_1)\boldsymbol{q}_1 - \cdots - (\boldsymbol{x}_j \vert \boldsymbol{q}_{j-1})\boldsymbol{q}_{j-1}$ e $\boldsymbol{q}_j = \frac{\boldsymbol{q}'_j}{\|\boldsymbol{q}'_j\|}$;
4º Passo: Voltar ao 2º passo.

2.4 Outros Resultados

Nesta secção, a última deste capítulo, apresentamos diversos resultados igualmente importantes para o desenvolvimento do trabalho futuro.

Desigualdades de Bonferroni

Consideremos n acontecimentos aleatórios $E_1, E_2, \ldots, E_n$. A probabilidade de que ocorram exactamente r acontecimentos ($r \leq n$), é dada por

$$P_{[r]} = \sum_{j=r}^{n} (-1)^{j-r} \binom{j}{j-r} S_j$$

onde, para $j > 0$,

$$S_j = \sum_{1 \leq \alpha_1 < \ldots < \alpha_j \leq n} Pr[\bigcap_{h=1}^{j} E_{\alpha_h}]$$

e $S_0 = 1$.

A probabilidade de que "pelo menos r"acontecimentos ocorram é

$$P_r = \sum_{j=r}^{n} P_{[j]} = \sum_{j=r}^{n} (-1)^{j-r} \binom{j-1}{j-r} S_j. \tag{2.23}$$

Bonferroni[20] provou (ver Bonferroni 1936) que

$$0 \leq P_{[0]} \leq 1$$

$$1 - S_1 \leq P_{[0]} \leq 1 - S_1 + S_2$$

e, de uma forma geral,

$$1 - S_1 + S_2 - \cdots - S_{2i-1} \leq P_{[0]} \leq 1 - S_1 + S_2 - \cdots + S_{2i} \tag{2.24}$$

para $i = 1, 2, \ldots, n/2$.

A demonstração destas desigualdades pode também ser vista em Galambos (1975 e 1977).

A desigualdade

$$1 - S_1 \leq P_{[0]}$$

é equivalente a

$$1 - \sum_{i=1}^{n} Pr[E_i] \leq Pr[\bigcap_{i=1}^{n} \overline{E_i}].$$

Substituindo o complemento $\overline{E_i}$ por A_i e, consequentemente, E_i por $\overline{A_i}$, obtemos

$$1 - \sum_{i=1}^{n} Pr[\overline{A_i}] \leq Pr[\bigcap_{i=1}^{n} A_i].$$

O conjunto de desigualdades (2.24) pode ser generalizado, obtendo-se

$$S_r - \left(\begin{array}{c} r+1 \\ 1 \end{array} \right) S_{r+1} \leq P_{[r]} \leq S_r$$

[20]Carlo Emilio Bonferroni (1892-1960) foi Professor Assistente do Politécnico de Turim e do Instituto Económico em Bari. Escreveu dois artigos abrangendo a sua desigualdade. O artigo de 1935 é dirigido a uma aplicação específica, o seguro de vida, enquanto que o artigo de 1936 é mais abstracto.

$$S_r - \binom{r+1}{1} S_{r+1} + \binom{r+2}{2} S_{r+2} - \binom{r+3}{3} S_{r+3} \leq$$

$$\leq P_{[r]} \leq S_r - \binom{r+1}{1} S_{r+1} + \binom{r+2}{2} S_{r+2}$$

$$\vdots$$

Similar conjunto de desigualdades pode ser obtido para P_r:

$$S_r - \binom{r}{1} S_{r+1} \leq P_r \leq S_r$$

$$S_r - \binom{r}{1} S_{r+1} + \binom{r+1}{2} S_{r+2} - \binom{r+2}{3} S_{r+3} \leq$$

$$\leq P_r \leq S_r - \binom{r}{1} S_{r+1} + \binom{r+1}{2} S_{r+2}$$

$$\vdots$$

Demonstrações destas desigualdades[21] podem ser encontradas em Frechet (1940) e Feller (1968).

Do ponto de vista estatístico, é a desigualdade de Bonferroni definida pela equação (2.23), que assume maior importância, pois tem grande utilidade na inferência estatística dado fornecer limites inferiores para a probabilidade de acontecimentos relevantes.

Fórmula de Faa di Bruno

A história da procura de uma expressão explícita para a derivada de ordem n da composição de funções é já bastante antiga.

Segundo Luckács (1955), a necessidade de uma expressão deste tipo foi mencionada pela primeira vez em 1810, no "Tratado de Cálculo"de Lacroix. Embora

[21] Meyer (1969) desenvolveu uma generalização das desigualdades de Bonferroni para o caso multivariado.

alguns casos particulares tenham sido apresentados, o primeiro a obter uma solução geral foi Faa di Bruno[22] em 1855.

Vejamos então a expressão obtida por Faa di Bruno.

Seja $g(x)$ uma função definida numa vizinhança do ponto x_0, admitindo neste ponto derivadas até à ordem n. Seja $f(y)$ uma outra função, definida numa vizinhança do ponto $y_0 = g(x_0)$, admitindo em y_0 derivadas até à ordem n.

A derivada de ordem n da função composição $h(x) = f[g(x)]$ no ponto x_0 é dada pela expressão

$$h_n = \sum_{k=1}^{n} f_k \sum_{p(n,k)} n! \prod_{i=1}^{n} \frac{g_i^{\lambda_i}}{(\lambda_i)!(i!)^{\lambda_i}}$$

onde

$$h_n = \tfrac{d^n}{dx^n} h(x_0), \qquad f_k = \tfrac{d^k}{dy^k} f(y_0), \qquad g_i = \tfrac{d^i}{dx^i} g(x_0)$$

e

$$p(n,k) = \{(\lambda_1, \ldots, \lambda_n) : \lambda_i \in N_0,\ \sum_{i=1}^{n} \lambda_i = k,\ \sum_{i=1}^{n} i\lambda_i = n\}$$

com N_0 um conjunto de inteiros não negativos.

O elemento $(\lambda_1, \ldots, \lambda_n) \in p(n,k)$ representa uma partição dum conjunto com n elementos em λ_1 classes de cardinalidade 1, λ_2 classes de cardinalidade 2,..., λ_n classes de cardinalidade n. O número destas partições é dado por S_n^k e denomina-se "número de Stirling de 2ª espécie"

$$S_n^k = \sum_{p(n,k)} n! \prod_{i=1}^{n} \frac{1}{(\lambda_i)!(i!)^{\lambda_i}}$$

tendo um papel fundamental na Teoria Combinatória (ver Constantine, 1987).

A fórmula de Faa di Bruno tem tido as mais diversas aplicações. Por exemplo, Luckács (1955) utilizou-a para relacionar os momentos e os cumulantes duma variável aleatória e para provar que uma população é normal se e só se a K-estatística[23] de ordem p ($\forall p > 1$) é independente da média da amostra.

[22] Francesco Faa di Bruno (1825-1888) nasceu em Alexandria, Itália. Começou a sua educação superior em 1841 na Academia Militar de Turim. No entanto, em 1853 abandonou o exército e dedicou-se ao estudo da Matemática. Viajou para Paris onde trabalhou sob a orientação de Cauchy. Em 1871 voltou à Itália onde se tornou professor na Universidade de Turim. O seu trabalho matemático mais famoso, publicado em 1876, foi na área das "formas binárias". Influenciado por Giovanni Bosco, Faa di Bruno foi ordenado Padre Católico, em Roma, no ano de 1876, tendo fundado uma ordem religiosa - Suore Minima di Nostra Signora del Suffragio.

[23] A K-estatística de ordem n de uma dada distribuição, K_n, é o único estimador centrado e simétrico do cumulante κ_n, isto é, $\mathrm{E}[K_n] = \kappa_n$. A K-estatística pode também ser definida através das somas $S_r \approx \sum_{i=1}^{n} X_i^r$, tendo-se

Mais recentemente Chen e Savits (1993) usaram-na para calcular momentos arbitrários dum processo de Poisson composto não homogéneo.

Desenvolvimentos mais aprofundados desta temática podem ser vistos, por exemplo, em Constantine e Savits (1996), Constantine (1987) e Luckacs (1955).

O Método da Bisseção

O método da bisseção é o mais simples dos métodos numéricos utilizados para obter a solução de uma equação não linear.

Consideremos uma função $f : [a, b] \to \mathbb{R}$, contínua, tal que $f(a)f(b) < 0$.

O teorema do valor intermédio garante-nos a existência de $\alpha \in]a, b[$ tal que $f(\alpha) = 0$.

Seja m o ponto médio do intervalo $[a, b]$. Se $f(m) = 0$, então m é uma raiz da equação. Se $f(a)f(m) < 0$, o teorema do valor intermédio garante a existência de uma raiz no intervalo $]a, m[$. Neste caso, tomamos $b = m$ e repetimos o procedimento.

Se $f(a)f(m) > 0$, temos que

$$f(a)f(b)f(a)f(m) < 0$$

ou seja

$$f(a)^2 f(b)f(m) < 0$$

e portanto

$$f(b)f(m) < 0$$

$$K_1 = \frac{S_1}{n}$$

$$K_2 = \frac{nS_2 - S_1^2}{n(n-1)}$$

$$K_3 = \frac{2S_1^3 - 3nS_1S_2 + n^2S_3}{n(n-1)(n-2)}$$

$$\vdots$$

(ver Fisher, 1928, Rose e Smith, 2002). Para uma amostra de dimensão n, as primeiras K-estatísticas são dadas por

$$K_1 = \mu$$

$$K_2 = \frac{n}{n-1} m_2$$

$$K_3 = \frac{n^2}{(n-1)(n-2)} m_3$$

$$\vdots$$

onde μ representa o valor médio da amostra, m_2 a variância da amostra e m_i o i-ésimo momento central da amostra (ver Kenney and Keeping, 1951 e 1962).

pelo que, novamente, pelo teorema do valor intermédio, podemos garantir a existência de uma raiz no intervalo $]m, b[$. Neste caso, tomamos $a = m$ e repetimos o procedimento.

Ao longo deste procedimento são criadas duas sequências numéricas constituídas pelos extremos dos intervalos $I_k = [a_{k-1}, b_{k-1}]$, cujo ponto médio é $m_k = \frac{b_{k-1}+a_{k-1}}{2}$.

Observe-se que, a cada passo do procedimento, se tem o intervalo $[a_k, b_k]$ que é dado por:

$$[a_{k-1}, m_k] \text{ se } f(a_{k-1})f(m_k) < 0$$

e

$$[m_k, b_{k-1}] \text{ se } f(a_{k-1})f(m_k) > 0.$$

Após n passos, a raiz estará contida no intervalo $[a_n, b_n]$, cujo comprimento é $b_n - a_n = \frac{b_0-a_0}{2^n}$. Então, a raiz α é tal que

$$|\alpha - m_{n+1}| < d_n = \frac{b_0 - a_0}{2^{n+1}}$$

que é uma estimativa para o erro absoluto dessa aproximação.
Se pretendermos que se verifique, para um dado ε,

$$0 < |\alpha - m_{n+1}| < \varepsilon$$

devemos escolher n tal que

$$n \geq \frac{ln(\frac{b_0-a_0}{\varepsilon})}{ln(2)} - 1.$$

Não podemos deixar de notar que as sequências (a_k) e (b_k), de extremos dos intervalos $[a_k, b_k]$, são convergentes, pois são monótonas e limitadas. Mais, prova-se que convergem para o mesmo limite.

De facto, como $b_k - a_k = \frac{b_0-a_0}{2^k}$, tem-se que

$$0 = \lim(b_k - a_k) = \lim b_k - \lim a_k$$

pelo que convergem para o mesmo limite L.

Como, em cada passo, se tem $f(a_k)f(b_k) < 0$, então

$$0 \geq \lim_{k\to\infty} f(a_k)f(b_k) = f(\lim_{k\to\infty} a_k)f(\lim_{k\to\infty} b_k) = f(L)^2 \geq 0$$

donde $f(L) = 0$.

Para terminar, falta apenas referir quais os critérios de paragem do algoritmo. Em geral o algoritmo deve parar quando se verifica uma ou ambas as condições:

$$\begin{cases} b_k - a_k < \varepsilon \\ \\ |f(m_k)| < \varepsilon. \end{cases}$$

Capítulo 3

Mínimos Quadrados Estruturados

3.1 Caso Geral

Comecemos pela noção de família parametrizada de modelos. Numa tal família, as matrizes dos modelos são identificadas por um parâmetro de estrutura $\boldsymbol{\theta}^w$, admitindo-se que o valor médio do vector das observações pertence ao espaço imagem $\Omega(\boldsymbol{\theta}^w_{\mathrm{V}})$ da matriz do modelo $\boldsymbol{X}(\boldsymbol{\theta}^w_{\mathrm{V}})$ quando $\boldsymbol{\theta}^w = \boldsymbol{\theta}^w_{\mathrm{V}}$, sendo este o verdadeiro valor do parâmetro de estrutura. As matrizes do modelo são todas do mesmo tipo, $n \times k$.

Sendo $\aleph = \{\boldsymbol{X}(\boldsymbol{\theta}^w) : \boldsymbol{\theta}^w \in \boldsymbol{\Gamma}\}$ a família de matrizes do modelo podemos definir duas relações de equivalência τ e τ°, em $\aleph$ e $\boldsymbol{\Gamma}$, respectivamente, por

$$\boldsymbol{X}(\boldsymbol{\theta}^w)\tau\boldsymbol{X}(\boldsymbol{\theta}'^w)$$

e

$$\boldsymbol{\theta}^w\tau^\circ\boldsymbol{\theta}'^w$$

quando

$$\Re(\boldsymbol{X}(\boldsymbol{\theta}^w)) = \Re(\boldsymbol{X}(\boldsymbol{\theta}'^w)),$$

sendo $\Re(\boldsymbol{M})$ o espaço imagem de $\boldsymbol{M}$.

Em geral, pode admitir-se que os vectores de estrutura identificam as matrizes do modelo tendo-se

$$\boldsymbol{X}(\boldsymbol{\theta}'^w) = \boldsymbol{X}(\boldsymbol{\theta}''^w)$$

se e só se

$$\boldsymbol{\theta}'^w = \boldsymbol{\theta}''^w.$$

Para realizar o ajustamento minimiza-se

$$S(\boldsymbol{\beta}^k, \boldsymbol{\theta}^w) = \|\boldsymbol{Y}^n - \boldsymbol{X}(\boldsymbol{\theta}^w)\boldsymbol{\beta}^k\|^2$$

para o que se pode utilizar um algoritmo tipo Zig-zag, no qual, a cada iteração se minimiza esta função primeiro em ordem a $\boldsymbol{\beta}^k$ e, de seguida, em ordem a $\boldsymbol{\theta}^w$.

Para iniciar o processo, é necessário escolher um valor inicial $\boldsymbol{\theta}^w_{(0)}$.

Se o modelo for homotópico, isto é se,

$$(c\boldsymbol{\theta}^w + d\mathbf{1}^w)\tau^\circ\boldsymbol{\theta}^w$$

quaisquer que sejam $\boldsymbol{\theta}^w \in \boldsymbol{\Gamma}$ e $c \neq 0$, pode-se no final de cada iteração j, $j = 1, 2, \ldots$, reajustar o vector $\widetilde{\boldsymbol{\theta}}^w_{(j)}$ obtido de forma a manter invariantes o mínimo e o máximo das suas componentes.

Representando por $\wedge\boldsymbol{v}^w$ e $\vee\boldsymbol{v}^w$ o mínimo e o máximo das componentes de $\boldsymbol{v}^w$, podemos, no caso dos modelos homotópicos, substituir $\boldsymbol{\theta}^w_{(1)}$, obtido minimizando $S(\boldsymbol{\theta}^w|\widetilde{\boldsymbol{\beta}}^w_{(1)})$, por $\widetilde{\boldsymbol{\theta}}^w_{(1)}$ tal que

$$\begin{cases} \widetilde{\boldsymbol{\theta}}^w_{(1)}\tau^\circ\boldsymbol{\theta}^w_{(1)} \\ \wedge\widetilde{\boldsymbol{\theta}}^w_{(1)} = \wedge\boldsymbol{\theta}^w_{(0)} \\ \vee\widetilde{\boldsymbol{\theta}}^w_{(1)} = \vee\boldsymbol{\theta}^w_{(0)}. \end{cases}$$

Como, com $\Omega(\boldsymbol{\theta}^w) = \Re(\boldsymbol{X}(\boldsymbol{\theta}^w))$, se tem

$$\min_{\boldsymbol{\beta}^k}\{S(\boldsymbol{\beta}^k|\boldsymbol{\theta}^w)\} = \|\boldsymbol{Y}^n_{\Omega(\boldsymbol{\theta}^w)^\perp}\|^2$$

onde $\boldsymbol{Y}^n_\nabla$ é a projecção ortogonal de $\boldsymbol{Y}^n$ sobre ∇ e $\Omega(\boldsymbol{\theta}^w)^\perp$ designa o complemento ortogonal de $\Omega(\boldsymbol{\theta}^w)$, ter-se-á

$$\min_{\boldsymbol{\beta}^k}\{S(\boldsymbol{\beta}^k|\widetilde{\boldsymbol{\theta}}^w_{(1)})\} = \min_{\boldsymbol{\beta}^k}\{S(\boldsymbol{\beta}^k|\boldsymbol{\theta}^w_{(1)})\}.$$

Pode, agora, iniciar-se uma segunda iteração com $\boldsymbol{\theta}^w = \widetilde{\boldsymbol{\theta}}^w_{(1)}$ e repetir o processo até que a soma de quadrados estabilize.

Dado proceder-se por minimizações sucessivas, os mínimos que se vão obtendo no final das iterações constituem uma sucessão decrescente; sendo limitada inferiormente por zero, é convergente.

Convém-nos considerar agora uma generalização directa da função objectivo em que se toma

$$S(\boldsymbol{\beta}^k|\widetilde{\boldsymbol{\theta}}^w) = (\boldsymbol{Y}^n - \boldsymbol{X}(\boldsymbol{\theta}^w)\boldsymbol{\beta}^k)^T\boldsymbol{C}(\boldsymbol{Y}^n - \boldsymbol{X}(\boldsymbol{\theta}^w)\boldsymbol{\beta}^k)$$

com $\boldsymbol{C}$ definida positiva[1]. Tem-se então

$$\boldsymbol{C} = (\boldsymbol{G}^T\boldsymbol{G})^{-1}$$

com

$$\boldsymbol{G} = \boldsymbol{D}^{-1/2}\boldsymbol{P}$$

[1]Uma matriz complexa de ordem n, $\boldsymbol{A}$, diz-se definida positiva se $\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} > 0$ para qualquer vector não nulo $\boldsymbol{x} \in \mathbb{C}$, onde $\boldsymbol{x}^*$ representa o transconjugado de $\boldsymbol{x}$, isto é, $\boldsymbol{x}^*{=}\overline{\boldsymbol{x}}^T$.

No caso de $\boldsymbol{A}$ ser uma matriz real, a equação reduz-se a $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0$.

onde $\boldsymbol{D}^{-1/2}$ é a matriz diagonal cujos elementos principais são as raízes quadradas dos inversos dos valores próprios de $\boldsymbol{C}$, $r_1, \ldots, r_n$, que são positivas, e $\boldsymbol{P}$ é a diagonalizadora[2] ortogonal de $\boldsymbol{C}$ cujos vectores linha são os vectores próprios de $\boldsymbol{C}$ associados aos valores próprios $r_1, \ldots, r_n$.

Então, com

$$\begin{cases} \boldsymbol{Y}'^n = \boldsymbol{G}\boldsymbol{Y}^n \\ \\ \boldsymbol{X}'(\boldsymbol{\theta}^w) = \boldsymbol{G}\boldsymbol{X}(\boldsymbol{\theta}^w), \end{cases}$$

a função objectivo reduz-se a

$$S(\boldsymbol{\beta}^k, \boldsymbol{\theta}^w) = \|\boldsymbol{Y}'^n - \boldsymbol{X}'(\boldsymbol{\theta}^w)\boldsymbol{\beta}^k\|^2$$

passando a aplicar-se as considerações precedentes.

Em particular, poder-se-á aplicar algoritmos tipo Zig-Zag.

3.2 Dois Factores Cruzados

Consideremos, agora, o caso particular em que se tem um modelo com dois factores $\boldsymbol{f}^m$ e $\boldsymbol{g}^s$, definindo-se

$$w = m + s$$

e

$$\boldsymbol{\theta}^w = \begin{pmatrix} \boldsymbol{f}^m \\ \\ \boldsymbol{g}^s \end{pmatrix}.$$

Representando por $\otimes$ o produto de Kronecker[3] de matrizes, a matriz do modelo tem a forma

$$\boldsymbol{X}(\boldsymbol{\theta}^w) = [\mathbf{1}^n | \boldsymbol{f}^m \otimes \mathbf{1}^s + \mathbf{1}^m \otimes \boldsymbol{g}^s],$$

e a matriz $\boldsymbol{C}$ é uma matriz diagonal cujos elementos principais são os pesos p_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, s$.

---

[2]Uma Matriz $\boldsymbol{P}$ diz-se ortogonal se $\boldsymbol{P}^T = \boldsymbol{P}^{-1}$.

Uma matriz $\boldsymbol{A}_{n\times n}$ é ortogonalmente diagonalizável se existir uma matriz ortogonal $\boldsymbol{P}_{n\times n}$ tal que

$$\boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P}$$

é uma matriz diagonal. A matriz $\boldsymbol{P}$ chama-se a diagonalizadora ortogonal de $\boldsymbol{A}$

[3]Dadas duas matrizes $\boldsymbol{A}_{m\times n}$ e $\boldsymbol{B}_{p\times q}$, o produto de Kronecker de $\boldsymbol{A}$ por $\boldsymbol{B}$, denotado por $\boldsymbol{A}\otimes\boldsymbol{B}$, é definido por

$$\boldsymbol{A}\otimes\boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{pmatrix}.$$

Dado um modelo da forma

$$y_{ij} = \beta_0 + \beta_1(f_i + g_j) \ , \ i = 1, \ldots, m, \ j = 1, \ldots, s$$

com pesos p_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, s$, para as observações da variável dependente, podemos representá-lo na forma geral através da introdução de índices

$$l = (i-1)s + j \ , \ i = 1, \ldots, m \ , \ j = 1, \ldots, s. \tag{3.1}$$

Nestes modelos, os parâmetros f_i, $i = 1, \ldots, m$, e g_j, $j = 1, \ldots, s$, correspondem aos níveis de dois factores. Por exemplo, exposição e susceptibilidade.

Para determinar $\boldsymbol{\theta}^w_{(0)}$ obtemos

$$\begin{cases} y_{i\bullet} = \frac{1}{s}\sum_{j=1}^{s} y_{ij} \ , \ i = 1, \ldots, m \\ \\ y_{\bullet j} = \frac{1}{m}\sum_{i=1}^{m} y_{ij} \ , \ j = 1, \ldots, s \end{cases}$$

e

$$y_{\bullet\bullet} = \frac{1}{ms}\sum_{i=1}^{m}\sum_{j=1}^{s} y_{ij}.$$

Calcula-se então

$$\theta_{ij}(0) = y_{i\bullet} + y_{\bullet j} - y_{\bullet\bullet} \ , \ i = 1, \ldots, m \ , \ j = 1, \ldots, s,$$

$a = \vee\boldsymbol{\theta}^w$ e $b = \wedge\boldsymbol{\theta}^w$ para, no fim de cada iteração $\imath$, $\imath = 1, 2, \ldots$, ajustar a escala.

A função objectivo pode agora, uma vez que o vector $\boldsymbol{\beta}$ tem apenas duas componentes, ser escrita na forma

$$S(\beta_0, \beta_1, \boldsymbol{f}^m, \boldsymbol{g}^s) = \sum_{i=1}^{m}\sum_{j=1}^{s} p_{ij}(y_{ij} - \beta_0 - \beta_1\theta_{ij}(0))^2.$$

Na primeira fase da iteração inicial, vamos minimizar esta função em ordem aos parâmetros β_0 e β_1.

Aplicando os resultados usuais da regressão linear, obtemos

$$\widetilde{\beta}_0(\imath) = Y_\circ - \widetilde{\beta}_1(\imath)\theta_{\bullet\bullet}(\imath)$$

com

$$\begin{cases} Y_\circ = \frac{1}{p^\oplus} \sum_{i=1}^{m} \sum_{j=1}^{s} p_{ij} y_{ij} \\ \theta_{\bullet\bullet}(\imath) = \frac{1}{p^\oplus} \sum_{i=1}^{m} \sum_{j=1}^{s} p_{ij} \theta_{ij}(\imath) \\ p^\oplus = \sum_{i=1}^{m} \sum_{j=1}^{s} p_{ij}, \end{cases}$$

bem como

$$\widetilde{\beta}_1(\imath) = \frac{s_{\theta y}(\imath)}{s_{\theta\Theta}(\imath)}$$

com

$$\begin{cases} \mathrm{s}_{\theta\Theta}(\imath) = S_{\theta\Theta}(\imath) - p^\oplus \theta^2_{\bullet\bullet}(\imath) \\ \mathrm{s}_{\theta y}(\imath) = S_{\theta y}(\imath) - p^\oplus \theta_{\bullet\bullet}(\imath) Y_\circ, \end{cases}$$

onde

$$\begin{cases} \mathrm{S}_{\theta\Theta}(\imath) = \sum_{i=1}^{m} \sum_{j=1}^{s} p_{ij} \theta^2_{ij}(\imath) \\ \mathrm{S}_{\theta y}(\imath) = \sum_{i=1}^{m} \sum_{j=1}^{s} p_{ij} \theta_{ij}(\imath) y_{ij}. \end{cases}$$

Minimiza-se de seguida, em ordem a $\boldsymbol{f}^m$ e $\boldsymbol{g}^s$, a função

$$S(\boldsymbol{f}^m, \boldsymbol{g}^s) = S(\boldsymbol{f}^m, \boldsymbol{g}^s | \widetilde{\beta}_0(\imath), \widetilde{\beta}_1(\imath)) = \sum_{i=1}^{m} \sum_{j=1}^{s} p_{ij} (y_{ij} - \widetilde{\beta}_0(\imath) - \widetilde{\beta}_1(\imath)(f_i + g_j))^2.$$

Como

$$\begin{cases} \frac{\partial S}{\partial f_i} = -2\widetilde{\beta}_1 \sum_{j=1}^{s} p_{ij} (y_{ij} - \widetilde{\beta}_0(\imath) - \widetilde{\beta}_1(\imath)(f_i - g_j)) \ , \ i = 1, \ldots, m \\ \frac{\partial S}{\partial g_j} = -2\widetilde{\beta}_1 \sum_{i=1}^{m} p_{ij} (y_{ij} - \widetilde{\beta}_0(\imath) - \widetilde{\beta}_1(\imath)(f_i - g_j)) \ , \ j = 1, \ldots, s \end{cases}$$

sendo $\boldsymbol{P} = [p_{ij}]$, ao anular-se estas derivadas obtém-se o sistema

$$\left(\begin{array}{c|c} \boldsymbol{D}_1 & \boldsymbol{P} \\ \hline \boldsymbol{P}^T & \boldsymbol{D}_2 \end{array}\right)\left(\begin{array}{c} \boldsymbol{f}^m \\ \hline \boldsymbol{g}^s \end{array}\right) = \boldsymbol{v}^w \tag{3.2}$$

onde

$$\begin{cases} \boldsymbol{D}_1 = \operatorname{diag}(\sum_{j=1}^{s} p_{1j}, \ldots, \sum_{j=1}^{s} p_{mj}) \\ \\ \boldsymbol{D}_2 = \operatorname{diag}(\sum_{i=1}^{m} p_{i1}, \ldots, \sum_{j=1}^{s} p_{is}) \end{cases}$$

tendo $\boldsymbol{v}^w$ as componentes

$$\begin{cases} v_i = \frac{1}{\widetilde{\beta}_1(\imath)} \sum_{j=1}^{s} p_{ij}(y_{ij} - \widetilde{\beta}_0(\imath)), \quad i = 1, \ldots, m \\ \\ v_{m+j} = \frac{1}{\widetilde{\beta}_1(\imath)} \sum_{i=1}^{m} p_{ij}(y_{ij} - \widetilde{\beta}_0(\imath)), \quad j = 1, \ldots, s. \end{cases}$$

Resolvendo-se o sistema (3.2) obtêm-se as soluções $\widetilde{\boldsymbol{f}}^m(\imath)$ e $\widetilde{\boldsymbol{g}}^s(\imath)$, donde resulta

$$\widetilde{\boldsymbol{\theta}}^w(\imath) = \widetilde{\boldsymbol{f}}^m(\imath) \otimes \mathbf{1}^s + \mathbf{1}^m \otimes \widetilde{\boldsymbol{g}}^s(\imath)$$

bem como

$$\begin{cases} \widetilde{a}(\imath) = \wedge\widetilde{\boldsymbol{\theta}}^w(\imath) \\ \\ \widetilde{b}(\imath) = \vee\widetilde{\boldsymbol{\theta}}^w(\imath). \end{cases}$$

Obtém-se agora o vector dado pelo ajustamento de escala

$$\check{\boldsymbol{\theta}}^w(\imath) = a\mathbf{1}^w + \frac{b-a}{\widetilde{b}(\imath) - \widetilde{a}(\imath)}(\widetilde{\boldsymbol{\theta}}^w(\imath) - \widetilde{a}(\imath)\mathbf{1}^w)$$

com

$$a = \wedge\boldsymbol{\theta}^w_{(0)} \quad \text{e} \quad b = \vee\boldsymbol{\theta}^w_{(0)}.$$

A soma de quadrados dos resíduos é, então,

$$S(\imath) = \sum_{i=1}^{m}\sum_{j=1}^{s} p_{ij}(y_{ij} - \widetilde{\beta}_0(\imath) - \widetilde{\beta}_1(\imath)\check{\theta}_{ij}(\imath))^2$$

onde $\check{\theta_{ij}}(\imath)$, $i = 1, \ldots, m$, $j = 1, \ldots, s$, são as componentes de $\check{\boldsymbol{\theta}}^w(\imath)$ uma vez determinados os índices l, de acordo com a equação (3.1).

Repetem-se os ciclos até $S(\imath)$, $\imath = 1, 2, \ldots$, estabilizar.

O procedimento que acabámos de descrever enquadra-se nos algoritmos tipo Zig-zag, ver Mexia *et al.* (2001), nome que se deve às minimizações alternadas.

Não podemos deixar de salientar que os resultados obtidos nesta secção podem generalizar-se para modelos com mais de dois factores.

3.3 Caso Normal

Admitamos que $\boldsymbol{y}^n$ é uma realização do vector $\boldsymbol{Y}^n$, normal, com vector médio $\boldsymbol{X}(\boldsymbol{\theta}^w)\boldsymbol{\beta}^k$ e matriz de covariância $\boldsymbol{C}$, onde $\boldsymbol{C}$ é conhecida e definida positiva. Ter-se-á então a densidade

$$n(\boldsymbol{y}^n|\boldsymbol{X}(\boldsymbol{\theta}^w)\boldsymbol{\beta}^k, \boldsymbol{C}) = \frac{e^{-\frac{1}{2}(\boldsymbol{y}^n - \boldsymbol{X}(\boldsymbol{\theta}^w)\boldsymbol{\beta}^k)^T\boldsymbol{C}^{-1}(\boldsymbol{y}^n - \boldsymbol{X}(\boldsymbol{\theta}^w)\boldsymbol{\beta}^k)}}{(2\pi)^{\frac{n}{2}}\sqrt{det(\boldsymbol{C})}}.$$

Assim, os estimadores que se obtêm maximizando a função objectivo serão estimadores de máxima verosimilhança. Este facto foi evidenciado em situações semelhantes por Pereira e Mexia (2004). Abrem-se, deste modo, grandes possibilidades de aplicação da Teoria Estatística. Em particular, poderão realizar-se testes de quocientes de verosimilhanças e aplicar versões assimptóticas do *Teorema de Wilks*. Estas possibilidades encontram-se exploradas, num contexto análogo no trabalho acima referido de Pereira e Mexia (2004). Vamos, no entanto, seguir uma via em que se utiliza a normalidade de uma forma mais adequada aos nossos objectivos.

Sendo $\widetilde{\boldsymbol{\beta}}^k$ e $\widetilde{\boldsymbol{\theta}}^w$ os parâmetros ajustados, admitamos que

$$\widetilde{\boldsymbol{\beta}}^k \sim N(\boldsymbol{\beta}^k, (\boldsymbol{X}^T(\boldsymbol{\theta}^w)\boldsymbol{X}(\boldsymbol{\theta}^w))^{-1})$$

e que

$$\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^w) \approx \boldsymbol{X}(\boldsymbol{\theta}^w),$$

o que nos permite trabalhar como se

$$\widetilde{\boldsymbol{\beta}}^k \overset{\bullet}{\sim} N(\boldsymbol{\beta}^k, (\boldsymbol{X}^T(\widetilde{\boldsymbol{\theta}}^w)\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^w))^{-1})$$

já que, ver Amemiya (1985) e Corte Real (2001), também se terá

$$(\boldsymbol{X}^T(\widetilde{\boldsymbol{\theta}}^w)\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^w))^{-1} \approx (\boldsymbol{X}^T(\boldsymbol{\theta}^w)\boldsymbol{X}(\boldsymbol{\theta}^w))^{-1}.$$

Tem-se então, sendo $\boldsymbol{\psi}^s = \boldsymbol{A}\boldsymbol{\beta}^k$, onde A tem característica s,

$$\widetilde{\boldsymbol{\psi}}^s = \boldsymbol{A}\widetilde{\boldsymbol{\beta}}^k \overset{\bullet}{\sim} N(\boldsymbol{\psi}^s, \boldsymbol{A}(\boldsymbol{X}^T(\widetilde{\boldsymbol{\theta}}^w)\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^w))^{-1}\boldsymbol{A}^T).$$

Portanto, para testar

$$H_0 : \boldsymbol{\psi}^s = \boldsymbol{\psi}_0^s$$

pode utilizar-se, ver Mexia (1990), a estatística

$$\Im(\boldsymbol{\psi}_0^s) = (\widetilde{\boldsymbol{\psi}}^s - \boldsymbol{\psi}_0^s)^T[\boldsymbol{A}(\boldsymbol{X}^T(\widetilde{\boldsymbol{\theta}}^w)\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^w))^{-1}\boldsymbol{A}^T]^{-1}(\widetilde{\boldsymbol{\psi}}^s - \boldsymbol{\psi}_0^s)$$

com (aproximadamente) distribuição χ^2, com s graus de liberdade e parâmetro de não centralidade

$$\delta = (\boldsymbol{\psi}^s - \boldsymbol{\psi}_0^s)^T[\boldsymbol{A}(\boldsymbol{X}^T(\widetilde{\boldsymbol{\theta}}^w)\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^w))^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{\psi}^s - \boldsymbol{\psi}_0^s)$$

que se anula se e só se H_0 se verificar.

O teste é, pois, (aproximadamente) não distorcido, ver Mexia (1990).

Por outro lado,

$$\Im' = \Im(\boldsymbol{\psi}^s)$$

terá (aproximadamente) distribuição χ^2 central com s graus de liberdade. Representando por $x_{1-\alpha,s}$ o quantil dessa distribuição para a probabilidade $(1-\alpha)$, ter-se-á (aproximadamente), ver Mexia (1990),

$$pr((\widetilde{\boldsymbol{\psi}}^s - \boldsymbol{\psi}_0^s)^T[\boldsymbol{A}(\boldsymbol{X}^T(\widetilde{\boldsymbol{\theta}}^w)\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^w))^{-1}\boldsymbol{A}^T]^{-1}(\widetilde{\boldsymbol{\psi}}^s - \boldsymbol{\psi}_0^s) \leq x_{1-\alpha,s}) = 1-\alpha.$$

Obtém-se assim um elipsóide de confiança de nível (aproximado) $1-\alpha$.

Conclui-se, deste modo, que o teste qui-quadrado de nível (aproximado) α não rejeita H_0 se e só se o elipsóide de confiança de nível (aproximado) $1-\alpha$ contiver $\boldsymbol{\psi}_0^s$. Assim, este teste goza de dualidade (aproximada).

Por outro lado um ponto está no interior dum elipsóide se e só se estiver entre todos os pares de planos paralelos tangentes ao elipsóide.

Dado que, para todo o $\boldsymbol{d}^s \neq \boldsymbol{0}^s$, existe um tal par de planos mostra-se, ver Scheffé (1959), que $\boldsymbol{\psi}^s$ está dentro do elipsóide de confiança quando e só quando

$$\bigcap_{\boldsymbol{d}^s}(|\boldsymbol{d}^{sT}\boldsymbol{\psi}^s - \boldsymbol{d}^{sT}\widetilde{\boldsymbol{\psi}}^s| \leq \sqrt{x_{1-\alpha,s}(\boldsymbol{d}^T\boldsymbol{A}(\boldsymbol{X}^T(\widetilde{\boldsymbol{\theta}}^w)\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^w))\boldsymbol{A}^T\boldsymbol{d})})$$

onde $\bigcap_{\boldsymbol{d}^s}$ indica que se consideram todos os vectores com s componentes.

Observe-se que a desigualdade correspondente a $\boldsymbol{d}^s = \boldsymbol{0}^s$ é da forma $0 \leq 0$, sendo portanto sempre satisfeita.

Ter-se-á pois, aproximadamente,

$$pr[\bigcap_{\boldsymbol{d}^s}(|\boldsymbol{d}^{sT}\boldsymbol{\psi}^s - \boldsymbol{d}^{sT}\widetilde{\boldsymbol{\psi}}^s| \leq \sqrt{x_{1-\alpha,s}(\boldsymbol{d}^T\boldsymbol{A}(\boldsymbol{X}^T(\widetilde{\boldsymbol{\theta}}^w)\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^w))\boldsymbol{A}^T\boldsymbol{d})})] = 1-\alpha.$$

Este resultado corresponde ao bem conhecido *Teorema de Scheffé*[4].

Quando

$$|\nu - \boldsymbol{d}^{sT}\widetilde{\boldsymbol{\psi}}^s| > \sqrt{x_{1-\alpha,s}(\boldsymbol{d}^T\boldsymbol{A}(\boldsymbol{X}^T(\widetilde{\boldsymbol{\theta}}^w)\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^w))\boldsymbol{A}^T\boldsymbol{d})}$$

a desigualdade correspondente a $\boldsymbol{d}^s$ não pode ser satisfeita com

$$\boldsymbol{d}^{sT}\boldsymbol{\psi}^s = \nu$$

pelo que, ao nível de aproximação que temos utilizado,

$$pr(\boldsymbol{d}^{sT}\boldsymbol{\psi}^s = \nu) < \alpha$$

dizendo-se que $\boldsymbol{d}^{sT}\boldsymbol{\psi}^s$ é significativamente diferente de ν a um nível aproximadamente igual a α.

A variante do *Teorema de Scheffé* apresentada dá-nos ainda intervalos de confiança simultâneos para as combinações lineares dos componentes de $\boldsymbol{\psi}^s$.

Observe-se que se considerarmos $\boldsymbol{A} = \boldsymbol{I}_k$ estamos a inferir sobre $\boldsymbol{\beta}^k$.

Se $s = 1$ e $\boldsymbol{A} = \boldsymbol{a}^T$, o elipsóide de confiança reduz-se a um simples intervalo de confiança. Não vale a pena, então, utilizar o teste χ^2, podendo testar-se

$$H_0 : \psi = \psi_0$$

mediante testes Z (aproximados) com estatística

$$z(\psi_0) = \frac{\widetilde{\psi} - \psi_0}{\sqrt{\boldsymbol{a}^T(\boldsymbol{X}^T(\widetilde{\boldsymbol{\theta}}^w)\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^w))^{-1}\boldsymbol{a}}}.$$

Estes testes podem ser unilaterais (direitos ou esquerdos) ou bilaterais.

Convém salientar que toda esta situação parece, pelo menos aparentemente, estar restringida ao caso uni-dimensional; vamos verificar na secção seguinte que não é verdade.

Por outro lado, ter-se-á no final da estimação

$$\widetilde{\boldsymbol{\theta}}^w = \boldsymbol{W}(\widetilde{\boldsymbol{\beta}}^k)(\boldsymbol{Y}^n - \boldsymbol{d}^n(\widetilde{\boldsymbol{\beta}}^k))$$

podendo, com o objectivo de estudar questões relativas a $\boldsymbol{\lambda}^l = \boldsymbol{B}\boldsymbol{\theta}^w$, e à semelhança do que se fez para $\widetilde{\boldsymbol{\beta}}^k$, admitir

$$\widetilde{\boldsymbol{\theta}}^w \overset{\bullet}{\sim} N(\boldsymbol{\theta}^w, \boldsymbol{W}(\widetilde{\boldsymbol{\beta}}^k)\boldsymbol{W}^T(\widetilde{\boldsymbol{\beta}}^k)).$$

[4]Teorema de Scheffé - Sejam $X, X_1, X_2, \ldots$ variáveis aleatórias contínuas definidas num espaço de probabilidade, cujas funções densidade de probabilidade são $f, f_1, f_2, \ldots$, respectivamente. Se $f_n \to f$ quase certamente, relativamente à medida de Lebesgue, então X_n converge em distribuição para X, isto é $X_n \xrightarrow{D} X$.

As expressões atrás apresentadas mantêm-se desde que se façam as substituições:

Antes	Depois
$\boldsymbol{A}$	$\boldsymbol{B}$
$\boldsymbol{X}^T(\widetilde{\boldsymbol{\theta}}^w)\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^w)$	$\boldsymbol{W}(\widetilde{\boldsymbol{\beta}}^k)\boldsymbol{W}^T(\widetilde{\boldsymbol{\beta}}^k)$
$\boldsymbol{a}^T$	$\boldsymbol{b}^T$
s	l

Os resultados normais que temos estado a considerar têm como pressuposto que as matrizes de covariância são conhecidas. Mexia, no trabalho (Mexia, 1990) em que os introduziu, designou estes modelos por "*error free*".

3.4 Testes χ^2 Selectivos

Nesta secção desenvolvemos testes de hipóteses para $\boldsymbol{\psi}^s$; no entanto, a teoria aplica-se a $\boldsymbol{\lambda}^l$ fazendo-se substituições perfeitamente semelhantes às indicadas no último parágrafo da secção anterior.

Consideremos, agora, coordenadas polares generalizadas $r, \gamma_1, \ldots, \gamma_{s-1}$. As coordenadas cartesianas, $x_1, \ldots, x_{s-1}$, podem ser obtidas a partir destas, ver Kendall (1961), utilizando as expressões

$$x_j = r\ell_i(\gamma_{s-1})$$

com

$$\begin{cases} \ell_1(\gamma_{s-1}) = \cos(\gamma_1)\cdots\cos(\gamma_{s-1}) \\ \ell_2(\gamma_{s-1}) = \cos(\gamma_1)\cdots\sin(\gamma_{s-1}) \\ \vdots \\ \ell_i(\gamma_{s-1}) = \cos(\gamma_1)\cdots\sin(\gamma_{s+1-i}) \\ \vdots \\ \ell_s(\gamma_{s-1}) = \sin(\gamma_1). \end{cases}$$

onde cos e sin indicam, respectivamente *coseno* e *seno*.

Para as coordenadas polares generalizadas têm-se os limites

$$\begin{cases} 0 \leq r \\ -\frac{\pi}{2} \leq \gamma_i \leq \frac{\pi}{2}\ ,\ i = 1, \ldots, s-2 \\ 0 \leq \gamma_{s-1} < 2\pi. \end{cases} \tag{3.3}$$

Seguindo um caminho semelhante ao de Mexia (1992), considerando

$$\boldsymbol{Z}^s \sim N(\boldsymbol{\eta}^s, \boldsymbol{C})$$

pretendemos obter a densidade conjunta de

$$\boldsymbol{\mathcal{V}} = (\boldsymbol{Z}^s - \boldsymbol{\eta}_0^s)^T \boldsymbol{W}^{-1} (\boldsymbol{Z}^s - \boldsymbol{\eta}_0^s)$$

com

$$\boldsymbol{W} = C$$

e do vector $\boldsymbol{\beth}^{s-1}$ dos ângulos ao centro de $(\boldsymbol{Z}^s - \boldsymbol{\eta}_0^s)$.

Observemos que

$$(\boldsymbol{Z}^s - \boldsymbol{\eta}^s)^T \boldsymbol{W}^{-1} (\boldsymbol{Z}^s - \boldsymbol{\eta}^s) = (\boldsymbol{Z}^s - \boldsymbol{\eta}_0^s + \boldsymbol{\eta}_0^s - \boldsymbol{\eta}^s)^T \boldsymbol{W}^{-1} (\boldsymbol{Z}^s - \boldsymbol{\eta}_0^s + \boldsymbol{\eta}_0^s - \boldsymbol{\eta}^s) =$$

$$= \boldsymbol{\mathcal{V}} + \boldsymbol{\delta} + 2(\boldsymbol{Z}^s - \boldsymbol{\eta}_0^s)^T \boldsymbol{W}^{-1} (\boldsymbol{\eta}_0^s - \boldsymbol{\eta}^s)$$

com

$$\boldsymbol{\delta} = (\boldsymbol{\eta}_0^s - \boldsymbol{\eta}^s)^T \boldsymbol{W}^{-1} (\boldsymbol{\eta}_0^s - \boldsymbol{\eta}^s),$$

pelo que a densidade de $\boldsymbol{Z}^s$ é dada por

$$n(\boldsymbol{z}^s | \boldsymbol{\eta}^s, \boldsymbol{W}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mathcal{V}} + \boldsymbol{\delta} + 2(\boldsymbol{Z}^s - \boldsymbol{\eta}_0^s)^T \boldsymbol{W}^{-1} (\boldsymbol{\eta}_0^s - \boldsymbol{\eta}^s))}}{(2\pi)^{s/2} \sqrt{det \boldsymbol{W}}}.$$

Considerando agora a transformação

$$\boldsymbol{Z}^s - \boldsymbol{\eta}_0^s = \boldsymbol{\mathcal{V}} \boldsymbol{\ell}^s(\boldsymbol{\gamma}^{s-1}) \tag{3.4}$$

e tomando-se

$$\begin{cases} k(\boldsymbol{\gamma}^{s-1}) = \boldsymbol{\ell}^s(\boldsymbol{\gamma}^{s-1})^T \boldsymbol{W}^{-1} \boldsymbol{\ell}^s(\boldsymbol{\gamma}^{s-1}) \\ a(\boldsymbol{\gamma}^{s-1}) = \boldsymbol{\ell}^s(\boldsymbol{\gamma}^{s-1})^T \boldsymbol{W}^{-1} (\boldsymbol{\eta}_0^s - \boldsymbol{\eta}^s) \end{cases}$$

tem-se

$$\boldsymbol{\mathcal{V}} = (\boldsymbol{Z}^s - \boldsymbol{\eta}_0^s)^T \boldsymbol{W}^{-1} (\boldsymbol{Z}^s - \boldsymbol{\eta}_0^s) = (\boldsymbol{\mathcal{V}} \boldsymbol{\ell}^s(\boldsymbol{\gamma}^{s-1}))^T \boldsymbol{W}^{-1} (\boldsymbol{\mathcal{V}} \boldsymbol{\ell}^s(\boldsymbol{\gamma}^{s-1})) =$$

$$= \boldsymbol{\mathcal{V}}^2 (\boldsymbol{\ell}^s(\boldsymbol{\gamma}^{s-1})^T \boldsymbol{W}^{-1} (\boldsymbol{\ell}^s(\boldsymbol{\gamma}^{s-1}) = \boldsymbol{\mathcal{V}}^2 k(\boldsymbol{\gamma}^{s-1})$$

bem como

$$(\boldsymbol{Z}^s - \boldsymbol{\eta}_0^s)^T \boldsymbol{W}^{-1}(\boldsymbol{\eta}_0^s - \boldsymbol{\eta}^s) = (\boldsymbol{\mathcal{V}}\boldsymbol{\ell}^s(\boldsymbol{\gamma}^{s-1}))^T \boldsymbol{W}^{-1}(\boldsymbol{\eta}_0^s - \boldsymbol{\eta}^s) =$$

$$= \boldsymbol{\mathcal{V}}(\boldsymbol{\ell}^s(\boldsymbol{\gamma}^{s-1}))^T \boldsymbol{W}^{-1}(\boldsymbol{\eta}_0^s - \boldsymbol{\eta}^s) = \boldsymbol{\mathcal{V}} a(\boldsymbol{\gamma}^{s-1}).$$

Como o Jacobiano da transformação para coordenadas polares generalizadas é, ver Kendall (1961, pág.17),

$$\mathcal{J}(v, \boldsymbol{\gamma}^{s-1}) = v^{s-1} h(\boldsymbol{\gamma}^{s-1})$$

com

$$h(\boldsymbol{\gamma}^{s-1}) = \prod_{i=1}^{s-2} c_i^{s-1-i}$$

e $c_i = cos(\gamma_i)$, $i = 1, \ldots, s-1$, a densidade conjunta de $\boldsymbol{\mathcal{V}}$ e $\boldsymbol{\beth}^{s-1}$ será

$$f(v, \boldsymbol{\gamma}^{s-1} | \boldsymbol{\eta}^s, \boldsymbol{W}) = v^{s-1} h(\boldsymbol{\gamma}^{s-1}) j(v, \boldsymbol{\gamma}^{s-1}) \frac{e^{-\frac{1}{2}(\boldsymbol{\mathcal{V}} + \boldsymbol{\delta} + 2\boldsymbol{\mathcal{V}} a(\boldsymbol{\gamma}^{s-1})}}{(2\pi)^{s/2}\sqrt{det\boldsymbol{W}}} \tag{3.5}$$

com $j(v)$ a função característica do domínio $\mathfrak{D}$.

Quando

$$H_0 : \boldsymbol{\eta}^s = \boldsymbol{\eta}_0^s$$

temos

$$\begin{cases} \delta = 0 \\ \\ a(\boldsymbol{\gamma}^{s-1}) = 0 \end{cases}$$

vindo

$$f(v, \boldsymbol{\gamma}^{s-1} | \boldsymbol{\eta}_0^s, \boldsymbol{W}) = \frac{v^{s-1} h(\boldsymbol{\gamma}^{s-1}) j(v, \boldsymbol{\gamma}^{s-1}) e^{-v/2}}{(2\pi)^{s/2}\sqrt{det\boldsymbol{W}}} = j^{\circ}(v) \frac{v^{s-1} e^{-v/2}}{2^s \Gamma(s)} m(\boldsymbol{\gamma}^{s-1})$$

com $j^{\circ}(v)$ a função característica do intervalo $[o, +\infty[$, Γ a Função Gama e

$$m(\boldsymbol{\gamma}^{s-1}) = \frac{j(v, \boldsymbol{\gamma}^{s-1}) h(\boldsymbol{\gamma}^{s-1})}{(2\pi)^{s/2}\sqrt{det\boldsymbol{W}}} \Gamma(s).$$

Conclui-se, assim, que $\boldsymbol{\mathcal{V}}$ é independente do vector $\boldsymbol{\beth}^{s-1}$ quando H_0 se verifica.

Suponhamos agora que se pretende construir um teste especialmente potente para certas situações especificadas a partir do vector

$$\beth^{s-1}(\boldsymbol{\eta}^s - \boldsymbol{\eta}_0^s)$$

dos ângulos ao centro de $(\boldsymbol{\eta}^s - \boldsymbol{\eta}_0^s)$.

Privilegiar-se-ão as situações tais que

$$(\boldsymbol{\eta}^s - \boldsymbol{\eta}_0^s) \in \mathfrak{D}_0,$$

com $\mathfrak{D}_0$ um sub-domínio do domínio $\mathfrak{D}$ de variação dos ângulos ao centro definidos em (3.3). É-se assim levado a rejeitar H_0 quando

$$\begin{cases} \widetilde{\beth} \in \mathfrak{D}_0 \\ \\ \boldsymbol{\mathcal{V}} > \mathbb{K}. \end{cases}$$

Atendendo à independência de $\boldsymbol{\mathcal{V}}$ e $\beth$ e a (3.5), o nível destes testes é dado por

$$\textit{nível}(\mathfrak{D}_0, \mathbb{K}) = pr(\boldsymbol{\mathcal{V}} > \mathbb{K}) pr(\widetilde{\beth} \in \mathfrak{D}_0) = (1 - F(\mathbb{K}|u)) \times \int \cdots \int_{\mathfrak{D}_0} m(\boldsymbol{u}^{s-1}) d_{u_1} \ldots d_{u_{s-1}}. \tag{3.6}$$

Suponhamos que as situações em que estamos interessados se caracterizam por

$$\|\boldsymbol{\eta}^s - \boldsymbol{\eta}_0^s\| \geq d$$

e por certas relações de ordem entre as componentes de $(\boldsymbol{\eta}^s - \boldsymbol{\eta}_0^s)$. Então, o integral que figura em (3.6) será o quociente das ordenações que satisfazem essas condições pelo número total de ordenações, $s!$. Com efeito, as condições nas ordenações exprimem-se através da escolha de $\mathfrak{D}_0$.

3.5 Validação do Modelo

Para além dos bem conhecidos erros de primeira e de segunda espécie existe, ver Tiago de Oliveira (1991), o erro de terceira espécie. Este caso verifica-se quando se escolhe um modelo errado.

Para nos protegermos deste erro, convém validar o modelo. Isto pode ser feito no final do ajustamento.

Com efeito, segundo a mesma perspectiva que atrás apresentámos, somos levados a verificar se

$$\boldsymbol{Y}^n = \boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^n)\widetilde{\boldsymbol{\beta}}^s + \boldsymbol{e}^n$$

com

$$\boldsymbol{e}^n \overset{\bullet}{\sim} N(\boldsymbol{\theta}^n, \boldsymbol{I}_n).$$

Sendo $\mathfrak{R}(\boldsymbol{M})$ o espaço imagem da matriz $\boldsymbol{M}$, aplicando o processo de ortogonalização de Gram-Schmidt aos vectores coluna de $\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^n)$, ver Bapat (2000), obtemos uma matriz $\boldsymbol{X}^{\ddagger}(\widetilde{\boldsymbol{\theta}}^n)$ com

$$\mathfrak{R}(\boldsymbol{X}^{\ddagger}(\widetilde{\boldsymbol{\theta}}^n)) = \mathfrak{R}(\boldsymbol{X}(\widetilde{\boldsymbol{\theta}}^n)).$$

Assim os vectores coluna de $\boldsymbol{X}^{\ddagger}(\widetilde{\boldsymbol{\theta}}^n)$ constituem uma base ortonormada para $\mathfrak{R}(\boldsymbol{X}^{\ddagger}(\widetilde{\boldsymbol{\theta}}^n))$ que pode ser completada por $\boldsymbol{\eta}_{s+1}^{m+1}, \ldots, \boldsymbol{\eta}_n^n$, de forma a obter uma base ortonormada para $\mathbb{R}^n$ (estes vectores constituem uma base ortonormada para $\mathfrak{R}(\boldsymbol{X}^{\ddagger}(\widetilde{\boldsymbol{\theta}}^n))^{\perp}$).

Sejam

$$\boldsymbol{A} = [\boldsymbol{\eta}_{s+1}^{m+1} \ldots \boldsymbol{\eta}_n^n]$$

e

$$\boldsymbol{e}_+^{n-s} = \boldsymbol{A}^T \boldsymbol{Y}^n.$$

Como, denotando por $\boldsymbol{O}_{r\times s}$ a matriz nula de tipo $r \times s$, se tem

$$\boldsymbol{A}^T \boldsymbol{X}(\boldsymbol{\theta}^n) = \boldsymbol{O}_{(n-s)\times s}$$

ter-se-á

$$\boldsymbol{e}_+^{n-s} = \boldsymbol{A}^T \boldsymbol{e}^n.$$

Assim, se o modelo se ajustar, teremos, pelo menos aproximadamente, ver Seber (1980, pg.8), que

$$\boldsymbol{e}_+^{n-s} \overset{\bullet}{\sim} N(\boldsymbol{0}^{n-s}, \boldsymbol{I}_{n-s}).$$

Para validar o modelo é-se assim levado a testar

$$H_0 : \boldsymbol{e}_+^{n-s} \overset{\bullet}{\sim} N(\boldsymbol{0}^{n-s}, \boldsymbol{I}_{n-s})$$

para o que se pode utilizar o *teste de Shapiro and Wilks*, ver Christensen (1987).

Em alternativa, pode-se utilizar o teste introduzido por Mexia (1989). Aliás, esta abordagem ao problema da validação segue o de Mexia (1989). Aproxima-se assim a Estatística da posição de *Popper* sobre as teorias científicas: "*uma teoria é científica se for falsificável*", isto é, se se puder provar que é falsa (ver Popper, 1995).

Aqui terá de se substituir "*teoria científica*" por "*modelo*" e "*falsa*" por "*não ajustado*". Aliás, uma teoria científica falsa corresponde, na ausência de contradições internas, a modelos que não se ajustam.

3.6 Logit

Para além da utilização de modelos "error free"vamos agora ver como aplicar a noção de verosimilhança nos modelos *logit*, uma vez introduzido o pressuposto de normalidade, isto é, onde

$$y_{ij} = logit(p_{ij}) = ln\frac{p_{ij}}{1-p_{ij}}$$

com p_{ij}, $i = 1, \ldots, n$ e $j = 1, \ldots, m$, a probabilidade de ocorrência de um dado acontecimento.

3.6.1 Logit e Verosimilhança

Consideremos n amostras, cada uma com m sub-amostras, de w elementos (indivíduos), nas quais existem x_{ij} indivíduos infectados, $i = 1, \ldots, n$ e $j = 1, \ldots, m$.

A função de verosimilhança pode ser escrita na forma

$$L_1(\beta_0, \beta_1, \boldsymbol{f}^{nm}|\boldsymbol{x}^{nm}) = \prod_{i=1}^{n}\prod_{j=1}^{m}\left(\begin{array}{c} w \\ x_{ij} \end{array} \right) \frac{e^{(\beta_0+\beta_1 f_{ij})x_{ij}}}{(1+e^{(\beta_0+\beta_1 f_{ij})})^w},$$

sendo o seu logaritmo

$$l_1(\beta_0, \beta_1, \boldsymbol{f}^{nm}|\boldsymbol{x}^{nm}) = \sum_{i=1}^{n}\sum_{j=1}^{m} ln\left(\begin{array}{c} w \\ x_{ij} \end{array} \right) + \sum_{i=1}^{n}\sum_{j=1}^{m}(\beta_0 + \beta_1 f_{ij})x_{ij} -$$

$$-w\sum_{i=1}^{n}\sum_{j=1}^{m} ln(1+e^{(\beta_0+\beta_1 f_{ij})})$$

com derivadas em ordem aos diferentes parâmetros

$$\begin{cases} \frac{\partial l_1}{\partial \beta_0} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} - w\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{e^{\beta_0+\beta_1 f_{ij}}}{1+e^{\beta_0+\beta_1 f_{ij}}} \\ \\ \frac{\partial l_1}{\partial \beta_1} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} f_{ij} - w\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(e^{\beta_0+\beta_1 f_{ij}})f_{ij}}{1+e^{\beta_0+\beta_1 f_{ij}}} \\ \\ \frac{\partial l_1}{\partial f_{ij}} = \beta_1 x_{ij} - w\beta_1\frac{e^{\beta_0+\beta_1 f_{ij}}}{1+e^{\beta_0+\beta_1 f_{ij}}} \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, m. \end{cases}$$

Na expressão de $L_1(\beta_0, \beta_1, \boldsymbol{f}^{nm}|\boldsymbol{x}^{nm})$, f_{ij} aparece integrada na "sub-função" $\beta_0 + \beta_1 f_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, portanto pode admitir-se que

$$\begin{cases} \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = 0 \\ \\ \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}^2 = 1. \end{cases} \tag{3.7}$$

Com

$$y_{ij} = \beta_0 + \beta_1 f_{ij} \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, m$$

obtemos, anulando as derivadas $\frac{\partial l_1}{\partial f_{ij}}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$,

$$x_{ij} = \frac{we^{y_{ij}}}{1 + e^{y_{ij}}}$$

isto é,

$$y_{ij} = \beta_0 + \beta_1 f_{ij} = ln\frac{x_{ij}}{w - x_{ij}} \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, m \tag{3.8}$$

vindo, devido à primeira das restrições da equação (3.7),

$$\widetilde{\beta}_0 = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} ln\frac{x_{ij}}{w - x_{ij}}.$$

Considerando

$$y_{\bullet\bullet} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}$$

podemos reescrever a equação (3.8) na forma

$$f_{ij} = \frac{1}{\beta_1}(y_{ij} - y_{\bullet\bullet}) \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, m,$$

vindo, devido à segunda restrição da equação (3.7),

$$\frac{1}{\beta_1^2}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - y_{\bullet\bullet})^2 = 1,$$

logo

$$\widetilde{\beta}_1^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - y_{\bullet\bullet})^2.$$

Por último, obtém-se

$$\widetilde{f}_{ij} = \frac{y_{ij} - y_{\bullet\bullet}}{\sqrt{\sum_{i'=1}^{n}\sum_{j'=1}^{m}(y_{i'j'} - y_{\bullet\bullet})^2}} \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, m.$$

Ter-se-á então

$$\widetilde{\beta}_0 + \widetilde{\beta}_1 \widetilde{f}_{ij} = \widetilde{y}_{ij} \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, m$$

pelo que

$$e^{\widetilde{\beta}_0 + \widetilde{\beta}_1 \widetilde{f}_{ij}} = \frac{x_{ij}}{w - x_{ij}} \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, m.$$

Assim, o máximo da verosimilhança será

$$\overline{L}_1 = \prod_{i=1}^{n}\prod_{j=1}^{m} \begin{pmatrix} w \\ x_{ij} \end{pmatrix} \frac{(\frac{x_{ij}}{w-x_{ij}})^{x_{ij}}}{(\frac{w}{w-x_{ij}})^{w}} = \prod_{i=1}^{n}\prod_{j=1}^{m} \begin{pmatrix} w \\ x_{ij} \end{pmatrix} \frac{(x_{ij})^{x_{ij}}(w - x_{ij})^{w-x_{ij}}}{w^w}.$$

Suponhamos que se pretende testar

$$H_{0,1} : f_{i1} = \cdots = f_{im} = f_i \ , \ i = 1, \ldots, n.$$

Com

$$x_i = \sum_{j=1}^{m} x_{ij}, \quad i = 1, \ldots, n$$

tem-se agora a verosimilhança "restringida pela hipótese"

$$L_2(\beta_0, \beta_1, \boldsymbol{f}_\circ^n | \boldsymbol{x}_\circ^n) = \prod_{i=1}^{n} \begin{pmatrix} mw \\ x_i \end{pmatrix} \frac{e^{(\beta_0 + \beta_1 f_i)x_i}}{(1 + e^{(\beta_0 + \beta_1 f_i)})^{mw}}$$

com $\boldsymbol{x}_\circ^n = (x_1, \ldots, x_n)^T$.

Tomemos

$$v_i = ln\frac{x_i}{mw - x_i} \ , \ i = 1, \ldots, n,$$

e

$$v_\circ = \frac{1}{n}\sum_{i=1}^{n} v_i.$$

Calculando o logaritmo de L_2, as suas derivadas e resolvendo o sistema que resulta da anulação das mesmas, obtemos os estimadores

$$\begin{cases} \widetilde{\beta}_{0,2} = v_\circ \\ \widetilde{\beta}_{1,2} = \sqrt{\sum_{i=1}^{n}(v_i - v_\circ)^2} \\ \widetilde{f}_i = \frac{v_i - v_\circ}{\sqrt{\sum_{i=1}^{n}(v_i - v_\circ)^2}}. \end{cases}$$

O máximo da verosimilhança, sob a restrição correspondente à hipótese testada, é

$$\overline{L}_2 = \prod_{i=1}^{n} \begin{pmatrix} mw \\ x_i \end{pmatrix} \frac{(x_i)^{x_i}(mw - x_i)^{mw - x_i}}{(mw)^{mw}}.$$

Atendendo ao *Teorema de Wilks*, quando $H_{0,1}$ se verifica a estatística

$$X_1 = -2ln\frac{L_2}{L_1}$$

tem (aproximadamente) distribuição qui-quadrado com $n(m-1)$ graus de liberdade, o que permite testar $H_{0,1}$.

Por outro lado, o efeito sobre os estimadores de β_0 e β_1, resultante da "agregação" das sub-amostras, é dado por

$$\begin{cases} \Delta_{\beta_0} = \widetilde{\beta}_{0,2} - \widetilde{\beta}_0 \\ \Delta_{\beta_1} = \widetilde{\beta}_{1,2} - \widetilde{\beta}_1. \end{cases}$$

Supondo que $H_{0,1}$ não foi rejeitada, podemos admitir a constância da exposição dentro de cada uma das n regiões. Admitamos ainda que nas regiões foram feitas novas medidas que, depois duma mudança de escala, são dadas por $f_1^\circ, \ldots, f_n^\circ$, verificando-se

$$\begin{cases} \sum_{i=1}^{n} f_i^\circ = 0 \\ \sum_{i=1}^{n} f_i^{\circ 2} = 1. \end{cases}$$

Interessa-nos testar

$$H_{0,2} : f_i = f_i^\circ \ , \ i = 1, \ldots, n.$$

Tem-se, agora, a verosimilhança

$$L_3(\beta_0, \beta_1 | \boldsymbol{x}_i^n) = \prod_{i=1}^{n} \left(\begin{array}{c} mw \\ x_i \end{array} \right) \frac{e^{(\beta_0+\beta_1 f_i^\circ)x_i}}{(1+e^{(\beta_0+\beta_1 f_i^\circ)})^{mw}}$$

cujo logaritmo é

$$l_3(\beta_0, \beta_1 | \boldsymbol{x}_i^n) = \sum_{i=1}^{n} ln \left(\begin{array}{c} mw \\ x_i \end{array} \right) + \sum_{i=1}^{n} (\beta_0 + \beta_1 f_i^\circ)x_i -$$

$$-mw \sum_{i=1}^{n} ln(1 + e^{\beta_0+\beta_1 f_i^\circ}).$$

Seguindo o mesmo raciocínio dos dois casos anteriores, isto é, calculando as derivadas da função l_3 em ordem aos parâmetros β_0 e β_1, obtemos

$$\begin{cases} \frac{\partial l_3}{\partial \beta_0} = \sum_{i=1}^{n} x_i - mw \sum_{i=1}^{n} \frac{e^{(\beta_0+\beta_1 f_i^\circ)}}{1 + e\beta_0 + \beta_1 f_i^\circ} \\ \\ \frac{\partial l_3}{\partial \beta_1} = \sum_{i=1}^{n} x_i f_i^\circ - mw \sum_{i=1}^{n} \frac{f_i^\circ e^{(\beta_0+\beta_1 f_i^\circ)}}{1 + e\beta_0 + \beta_1 f_i^\circ}. \end{cases}$$

Anulando $\frac{\partial l_3}{\partial \beta_0}$ e $\frac{\partial l_3}{\partial \beta_1}$, obtemos o sistema

$$\begin{cases} \sum_{i=1}^{n} x_i - mw \sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 f_i^\circ}}{1 + e^{\beta_0+\beta_1 f_i^\circ}} = 0 \\ \\ \sum_{i=1}^{n} x_i f_i^\circ - mw \sum_{i=1}^{n} \frac{f_i^\circ e^{\beta_0+\beta_1 f_i^\circ}}{1 + e^{\beta_0+\beta_1 f_i^\circ}} = 0. \end{cases}$$

A resolução deste sistema, que obviamente terá que ser feita numericamente, permitirá obter os estimadores $\widetilde{\beta}_{0,3}$ e $\widetilde{\beta}_{1,3}$.

O máximo da verosimilhança L_3 será, então,

$$\overline{L}_3 = \prod_{i=1}^{n} \left(\begin{array}{c} mw \\ x_i \end{array} \right) \frac{e^{(\widetilde{\beta}_{0,3}+\widetilde{\beta}_{1,3} f_i^\circ)x_i}}{(1+e^{(\widetilde{\beta}_{0,3}+\widetilde{\beta}_{1,3} f_i^\circ)})^{mw}}.$$

Atendendo novamente ao *Teorema de Wilks*, quando $H_{0,2}$ se verifica a estatística

$$X_2 = -2ln\frac{L_3}{L_2}$$

tem (aproximadamente) distribuição qui-quadrado com n graus de liberdade.

Observemos ainda que, considerando

$$\begin{cases} \overline{L}_1 = \max\{L_1\} \\ \\ \overline{L}_2 = \max\{L_2\} \\ \\ \overline{L}_3 = \max\{L_3\} \end{cases}$$

têm-se as perdas

$$\begin{cases} \overline{L}_2 - \overline{L}_1 \rightarrow \text{perda devido à passagem de } L_1 \text{ para } L_2 \\ \\ \overline{L}_3 - \overline{L}_2 \rightarrow \text{perda devido à passagem de } L_2 \text{ para } L_3 \ . \end{cases}$$

O teste de Wilks permite-nos verificar se estas perdas são ou não significativas.

Generalização

Consideremos, agora, uma generalização em que se tem n amostras, com m_i sub-amostras, de w_{ij} elementos (indivíduos) e x_{ij} indivíduos infectados, com $i = 1, \ldots, n$ e $j = 1, \ldots, m_i$. Consideremos ainda

$$\begin{cases} \overline{m} = \sum_{i=1}^{n} m_i \\ \\ \overline{w} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}. \end{cases}$$

Desta forma, tem-se

$$L_1(\beta_0, \beta_1, \boldsymbol{f}^{\overline{m}} | \boldsymbol{x}^{\overline{m}}) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} \begin{pmatrix} w_{ij} \\ x_{ij} \end{pmatrix} \frac{e^{(\beta_0+\beta_1 f_{ij})x_{ij}}}{(1+e^{(\beta_0+\beta_1 f_{ij})})^{w_{ij}}}$$

cujo logaritmo é

$$l_1(\beta_0, \beta_1, \boldsymbol{f}^{\overline{m}} | \boldsymbol{x}^{\overline{m}}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} ln \begin{pmatrix} w_{ij} \\ x_{ij} \end{pmatrix} + \sum_{i=1}^{n} \sum_{j=1}^{m_i} (\beta_0 + \beta_1 f_{ij}) x_{ij} -$$

$$- \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij} ln(1 + e^{(\beta_0+\beta_1 f_{ij})})$$

com derivadas em ordem aos diferentes parâmetros

$$\begin{cases} \frac{\partial l_1}{\partial \beta_0} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} x_{ij} - \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\frac{e^{\beta_0+\beta_1 f_{ij}}}{1+e^{\beta_0+\beta_1 f_{ij}}} \\ \\ \frac{\partial l_1}{\partial \beta_1} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} x_{ij} f_{ij} - \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\frac{(e^{\beta_0+\beta_1 f_{ij}})f_{ij}}{1+e^{\beta_0+\beta_1 f_{ij}}} \\ \\ \frac{\partial l_1}{\partial f_{ij}} = \beta_1 x_{ij} - w_{ij}\frac{e^{\beta_0+\beta_1 f_{ij}}}{1+e^{\beta_0+\beta_1 f_{ij}}}\beta_1 \ , \ i=1,\ldots,n \ , \ j=1,\ldots,m_i. \end{cases} \tag{3.9}$$

Dado que, nas expressões de L_1 e l_1, f_{ij} aparece integrada na "sub-função" $\beta_0 + \beta_1 f_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, pode admitir-se que

$$\begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij} f_{ij} = 0 \\ \\ \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij} f_{ij}^2 = 1. \end{cases} \tag{3.10}$$

Considerando

$$y_{ij} = \beta_0 + \beta_1 f_{ij} \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, m_i$$

anulando as derivadas $\frac{\partial l_1}{\partial f_{ij}}$, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, obtemos

$$x_{ij} = \frac{w_{ij} e^{y_{ij}}}{1 + e^{y_{ij}}}$$

vindo

$$y_{ij} = \beta_0 + \beta_1 f_{ij} = ln\frac{x_{ij}}{w_{ij} - x_{ij}} \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, m_i \tag{3.11}$$

e, devido à primeira das restrições da equação (3.10),

$$\widetilde{\beta}_0 = \frac{1}{w}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij} ln\frac{x_{ij}}{w_{ij} - x_{ij}}.$$

Tomando

$$y_{\bullet\bullet} = \frac{1}{w}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij} y_{ij}$$

tem-se

$$\widetilde{\beta}_0 = y_{\bullet\bullet}$$

pelo que a equação (3.11) pode escrever-se na forma

$$f_{ij} = \frac{1}{\beta_1}(y_{ij} - y_{\bullet\bullet}), \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i.$$

Da segunda das restrições da equação (3.10) resulta que

$$\frac{1}{\beta_1^2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}(y_{ij} - y_{\bullet\bullet})^2 = 1$$

e, consequentemente,

$$\widetilde{\beta}_1 = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}(y_{ij} - y_{\bullet\bullet})^2}.$$

O que é coerente com o usual pressuposto de que $\beta_1 > 0$, que traduz o facto da probabilidade de doença aumentar com a exposição.

Por último, obtemos

$$\widetilde{f}_{ij} = \frac{y_{ij} - y_{\bullet\bullet}}{\sum_{i'=1}^{n} \sum_{j'=1}^{m_{i'}} w_{i'j'}(y_{i'j'} - y_{\bullet\bullet})^2} \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, m_i.$$

Ter-se-á então

$$\widetilde{\beta}_0 + \widetilde{\beta}_1 \widetilde{f}_{ij} = \widetilde{y}_{ij} \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, m_i$$

vindo

$$e^{\widetilde{\beta}_0 + \widetilde{\beta}_1 \widetilde{f}_{ij}} = \frac{x_{ij}}{w_{ij} - x_{ij}} \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, m_i.$$

Assim, o máximo da verosimilhança será

$$\overline{L}_1 = \prod_{i=1}^{n} \prod_{j=1}^{m_i} \begin{pmatrix} w_{ij} \\ x_{ij} \end{pmatrix} \frac{(\frac{x_{ij}}{w_{ij} - x_{ij}})^{x_{ij}}}{(\frac{w_{ij}}{w_{ij} - x_{ij}})^{w_{ij}}} = \prod_{i=1}^{n} \prod_{j=1}^{m_i} \begin{pmatrix} w_{ij} \\ x_{ij} \end{pmatrix} \frac{(x_{ij})^{x_{ij}} (w_{ij} - x_{ij})^{w_{ij} - x_{ij}}}{w_{ij}^{w_{ij}}}.$$

Com o objectivo de obter a matriz de covariância dos nossos estimadores, começamos por calcular a matriz Hessiana de $l_1(\beta_0, \beta_1, \boldsymbol{f}^{\overline{m}} | \boldsymbol{x}^{\overline{m}})$.

A partir das expressões (3.9) é fácil obter as derivadas de 2ª ordem:

$$\begin{cases} \frac{\partial^2 l_1}{\partial \beta_0^2} = -\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\frac{e^{\beta_0+\beta_1 f_{ij}}}{(1+e^{\beta_0+\beta_1 f_{ij}})^2} \\ \frac{\partial^2 l_1}{\partial \beta_0 \partial \beta_1} = -\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\frac{e^{\beta_0+\beta_1 f_{ij}}}{(1+e^{\beta_0+\beta_1 f_{ij}})^2} f_{ij} \\ \frac{\partial^2 l_1}{\partial \beta_0 \partial f_{ij}} = -w_{ij}\frac{e^{\beta_0+\beta_1 f_{ij}}}{(1+e^{\beta_0+\beta_1 f_{ij}})^2}\beta_1 \\ \frac{\partial^2 l_1}{\partial \beta_1^2} = -\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\frac{e^{\beta_0+\beta_1 f_{ij}}}{(1+e^{\beta_0+\beta_1 f_{ij}})^2} f_{ij}^2 \\ \frac{\partial^2 l_1}{\partial \beta_1 \partial f_{ij}} = x_{ij} - w_{ij}[\frac{e^{\beta_0+\beta_1 f_{ij}}}{(1+e^{\beta_0+\beta_1 f_{ij}})^2}\beta_1 f_{ij} + \frac{e^{\beta_0+\beta_1 f_{ij}}}{1+e^{\beta_0+\beta_1 f_{ij}}}] \\ \frac{\partial^2 l_1}{\partial f_{ij} \partial f_{i'j'}} = -\delta_{ii'}\delta_{jj'} w_{ij}\frac{e^{\beta_0+\beta_1 f_{ij}}}{(1+e^{\beta_0+\beta_1 f_{ij}})^2}\beta_1^2 \end{cases}$$

com $i, i' = 1, \ldots, n$, $j = 1, \ldots, m_i$, $j' = 1, \ldots, m_{i'}$ e onde δ_{kt} representa o símbolo de Kronecker, $\delta_{kt} = 1$ se $k = t$ e $\delta_{kt} = 0$ se $k \neq t$.

Quando substituímos os parâmetros pelos seus estimadores, obtemos

$$\begin{cases} \frac{\partial^2 l_1}{\partial \beta_0^2} = -\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\frac{(x_{ij})^{x_{ij}}(w_{ij}-x_{ij})^{2w_{ij}-x_{ij}}}{(w_{ij})^{2w_{ij}}} \\ \frac{\partial^2 l_1}{\partial \beta_0 \partial \beta_1} = -\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\frac{(x_{ij})^{x_{ij}}(w_{ij}-x_{ij})^{2w_{ij}-x_{ij}}}{(w_{ij})^{2w_{ij}}} f_{ij} \\ \frac{\partial^2 l_1}{\partial \beta_0 \partial f_{ij}} = -w_{ij}\frac{(x_{ij})^{x_{ij}}(w_{ij}-x_{ij})^{2w_{ij}-x_{ij}}}{(w_{ij})^{2w_{ij}}}\beta_1 \\ \frac{\partial^2 l_1}{\partial \beta_1^2} = -\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\frac{(x_{ij})^{x_{ij}}(w_{ij}-x_{ij})^{2w_{ij}-x_{ij}}}{(w_{ij})^{2w_{ij}}} f_{ij}^2 \\ \frac{\partial^2 l_1}{\partial \beta_1 \partial f_{ij}} = x_{ij} - w_{ij}[\frac{(x_{ij})^{x_{ij}}(w_{ij}-x_{ij})^{2w_{ij}-x_{ij}}}{(w_{ij})^{2w_{ij}}}\beta_1 f_{ij} + \frac{(x_{ij})^{x_{ij}}(w_{ij}-x_{ij})^{w_{ij}-x_{ij}}}{(w_{ij})^{w_{ij}}}] \\ \frac{\partial^2 l_1}{\partial f_{ij} \partial f_{i'j'}} = -\delta_{ii'}\delta_{jj'} w_{ij}\frac{(x_{ij})^{x_{ij}}(w_{ij}-x_{ij})^{2w_{ij}-x_{ij}}}{(w_{ij})^{2w_{ij}}}\beta_1^2 \end{cases}$$

com $i, i' = 1, \ldots, n$, $j = 1, \ldots, m_i$ e $j' = 1, \ldots, m_{i'}$.

Suponhamos que se pretende testar

$$H_{0,1} : f_{i1} = \cdots = f_{im_i} = f_i \ , \ i = 1, \ldots, n.$$

Com

$$\begin{cases} w_i = \sum_{j=1}^{m_i} w_{ij} \ , \ i = 1, \ldots, n \\ \\ x_i = \sum_{j=1}^{m_i} x_{ij} \ , \ i = 1, \ldots, n \end{cases}$$

tem-se agora a verosimilhança "restringida pela hipótese"

$$L_2(\beta_0, \beta_1, \boldsymbol{f}_\circ^n | \boldsymbol{x}_\circ^n) = \prod_{i=1}^{n} \left(\begin{array}{c} w_i \\ x_i \end{array} \right) \frac{e^{(\beta_0+\beta_1 f_i)x_i}}{(1+e^{(\beta_0+\beta_1 f_i)})^{w_i}}.$$

Com

$$v_i = ln \frac{x_i}{w_i - x_i} \ , \ i = 1, \ldots, n$$

e

$$v_\circ = \frac{1}{n} \sum_{i=1}^{n} w_i v_i$$

obtemos os seguintes estimadores, que resultam da anulação das derivadas do logaritmo de L_2,

$$\begin{cases} \widetilde{\beta}_{0,2} = v_\circ \\ \\ \widetilde{\beta}_{1,2} = \sqrt{\sum_{i=1}^{n} w_i (v_i - v_\circ)^2} \\ \\ \widetilde{f}_{i,2} = \frac{v_i - v_\circ}{\sqrt{\sum_{i=1}^{n} w_i (v_i - v_\circ)^2}}. \end{cases}$$

Assim, o máximo da verosimilhança condicionada será

$$\overline{L}_2 = \prod_{i=1}^{n} \left(\begin{array}{c} w_i \\ x_i \end{array} \right) \frac{(x_i)^{x_i} (w_i - x_i)^{w_i - x_i}}{(w_i)^{w_i}}.$$

Atendendo ao *Teorema de Wilks*, quando $H_{0,1}$ se verifica a estatística

$$X_1 = -2ln \frac{L_2}{L_1}$$

tem (aproximadamente) distribuição qui-quadrado com $(\overline{m}-n)$ graus de liberdade.

Por outro lado, o efeito sobre os estimadores de β_0 e β_1, resultante da "agregação" das sub-amostras, é dado por

$$\begin{cases} \Delta_{\beta_0} = \widetilde{\beta}_{0,2} - \widetilde{\beta}_0 \\ \\ \Delta_{\beta_1} = \widetilde{\beta}_{1,2} - \widetilde{\beta}_1. \end{cases}$$

Supondo que $H_{0,1}$ não foi rejeitada, podemos admitir a constância da exposição dentro de cada uma das n regiões seleccionadas. Admitamos ainda que nas regiões foram feitas novas medidas que, depois duma mudança de escala, são dadas por $f_1^\circ, \ldots, f_n^\circ$, verificando-se

$$\begin{cases} \sum_{i=1}^{n} f_i^\circ = 0 \\ \\ \sum_{i=1}^{n} f_i^{\circ 2} = 1. \end{cases}$$

Interessa-nos agora testar

$$H_{0,2} : f_i = f_i^\circ \ , \ i = 1, \ldots, n.$$

Tem-se a verosimilhança

$$L_3(\beta_0, \beta_1 | \boldsymbol{x}_i^n) = \prod_{i=1}^{n} \left(\begin{array}{c} w_i \\ x_i \end{array} \right) \frac{e^{(\beta_0+\beta_1 f_i^\circ)x_i}}{(1+e^{(\beta_0+\beta_1 f_i^\circ)})^{w_i}}$$

cujo logaritmo é

$$l_3(\beta_0, \beta_1 | \boldsymbol{x}_i^n) = \sum_{i=1}^{n} ln \left(\begin{array}{c} w_i \\ x_i \end{array} \right) + \sum_{i=1}^{n} (\beta_0 + \beta_1 f_i^\circ)x_i - \sum_{i=1}^{n} w_i ln(1 + e^{\beta_0+\beta_1 f_i^\circ}).$$

Seguindo o mesmo raciocínio dos dois casos anteriores, isto é, calculando as derivadas da função l_3 em ordem aos parâmetros β_0 e β_1, obtemos

$$\begin{cases} \frac{\partial l_3}{\partial \beta_0} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} w_i \frac{e^{(\beta_0+\beta_1 f_i^\circ)}}{1 + e\beta_0 + \beta_1 f_i^\circ} \\ \\ \frac{\partial l_3}{\partial \beta_1} = \sum_{i=1}^{n} x_i f_i^\circ - \sum_{i=1}^{n} w_i \frac{f_i^\circ e^{(\beta_0+\beta_1 f_i^\circ)}}{1 + e\beta_0 + \beta_1 f_i^\circ}. \end{cases}$$

Anulando $\frac{\partial l_3}{\partial \beta_0}$ e $\frac{\partial l_3}{\partial \beta_1}$ obtemos o sistema

$$\begin{cases} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} w_i \frac{e^{\beta_0+\beta_1 f_i^\circ}}{1+e^{\beta_0+\beta_1 f_i^\circ}} = 0 \\ \sum_{i=1}^{n} x_i f_i^\circ - \sum_{i=1}^{n} w_i \frac{f_i^\circ e^{\beta_0+\beta_1 f_i^\circ}}{1+e^{\beta_0+\beta_1 f_i^\circ}} = 0. \end{cases}$$

A resolução deste sistema, que se obterá recorrendo a métodos numéricos, permitirá obter os estimadores $\widetilde{\beta}_{0,3}$ e $\widetilde{\beta}_{1,3}$.

O máximo da verosimilhança L_3 será, então,

$$\overline{L}_3 = \prod_{i=1}^{n} \left(\begin{array}{c} w_i \\ x_i \end{array} \right) \frac{e^{(\widetilde{\beta}_{0,3}+\widetilde{\beta}_{1,3} f_i^\circ) x_i}}{(1+e^{(\widetilde{\beta}_{0,3}+\widetilde{\beta}_{1,3} f_i^\circ)})^{w_i}}.$$

Atendendo novamente ao *Teorema de Wilks*, quando $H_{0,2}$ se verifica a estatística

$$X_2 = -2ln\frac{L_3}{L_2}$$

tem (aproximadamente) distribuição qui-quadrado com n graus de liberdade.

Observemos ainda que, considerando

$$\begin{cases} \overline{L}_1 = \max\{L_1\} \\ \overline{L}_2 = \max\{L_2\} \\ \overline{L}_3 = \max\{L_3\} \end{cases}$$

têm-se as perdas

$$\begin{cases} \overline{L}_2 - \overline{L}_1 \rightarrow \text{perda devido à passagem de } L_1 \text{ para } L_2 \\ \overline{L}_3 - \overline{L}_2 \rightarrow \text{perda devido à passagem de } L_2 \text{ para } L_3 \ . \end{cases}$$

O teste de Wilks permite-nos verificar se estas perdas são ou não significativas.

3.6.2 Logit, Verosimilhança e Aditividade

Admitamos que existem s sub-populações com susceptibilidades g_j, $j = 1, \ldots, s$, e que, quando se consideram os níveis de exposição f_i, $i = 1, \ldots, n$, os logits das probabilidades de doença, p_{ij}, são dados por

$$y_{ij} = logit(p_{ij}) = \beta_0 + \beta_1(f_i + g_j) \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, s$$

havendo pois aditividade entre susceptibilidade e exposição.

Tomando-se

$$f_\bullet = \tfrac{1}{n}\sum_{i=1}^{n} f_i \qquad \text{e} \qquad g_\bullet = \tfrac{1}{s}\sum_{j=1}^{s} g_j$$

ter-se-á

$$y_{ij} = y_{\bullet\bullet} + \alpha_i + \gamma_j \ , \ i = 1, \ldots, n \ , \ j = 1, \ldots, s$$

com

$$\begin{cases} y_{\bullet\bullet} = \beta_0 + \beta_1(f_\bullet + g_\bullet) \\ \alpha_i = \beta_1(f_i - f_\bullet) \ , \ i = 1, \ldots, n \\ \gamma_j = \beta_1(g_j - g_\bullet) \ , \ j = 1, \ldots, s. \end{cases}$$

É fácil de verificar que

$$\sum_{i=1}^{n} \alpha_i = \sum_{j=1}^{s} \gamma_j = 0.$$

A matriz do modelo

$$\boldsymbol{X} = [\mathbf{1}^{ns}\boldsymbol{I}_n \otimes \mathbf{1}^h\mathbf{1}^n \otimes \boldsymbol{I}_s]$$

corresponde ao delineamento completo, podendo os pares (i, j) ser ordenados de acordo com o índice

$$h = (i-1)s + j.$$

Atendendo-se às igualdades

$$\begin{cases} \alpha_n = -\sum_{i=1}^{n-1} \alpha_i \\ \gamma_s = -\sum_{j=1}^{s-1} \gamma_j \end{cases}$$

é possível eliminar os parâmetros redundantes, condensando a matriz $\boldsymbol{X}$, de tipo $ns \times (n+s+1)$, obtendo-se uma nova matriz do modelo, $\boldsymbol{X}^\circ$, de tipo $(ns) \times w$, com $w = n + s - 1$, a que correspondem os parâmetros $y_{\bullet\bullet}$, $\boldsymbol{\alpha}'^{(n-1)} = (\alpha_1, \ldots, \alpha_{n-1})^T$ e $\boldsymbol{\gamma}'^{(s-1)} = (\gamma_1, \ldots, \gamma_{s-1})^T$.

A aditividade permite-nos colher amostras apenas para parte dos pares (i, j).

Seja $\boldsymbol{X}_r$ a sub-matriz formada pelas n_r linhas de $\boldsymbol{X}^\circ$ correspondentes aos pares escolhidos. Sendo

$$\text{car}(\boldsymbol{X}_r) = n + s - 1$$

os parâmetros $y_{\bullet\bullet}$, $\boldsymbol{\alpha}'^{(n-1)}$ e $\boldsymbol{\gamma}'^{(s-1)}$ são estimáveis.

Seja C o conjunto de pares (i, j) escolhidos. Utilizando amostras de dimensão m, teremos os estimadores para as probabilidades de doença

$$\widehat{p}_{ij} = \frac{x_{ij}}{m} \ , \ (i, j) \in C.$$

A partir destes, obtemos os estimadores

$$\widehat{y}_{ij} = logit(\widehat{p}_{ij}) \ , \ (i, j) \in C$$

com o inverso das variâncias

$$v_{ij} \approx \frac{m}{p_{ij}} \ , \ (i, j) \in C$$

que podem ser estimadas por

$$\widehat{v}_{ij} = \frac{m}{x_{ij}} \ , \ (i, j) \in C.$$

No que segue, substituiremos as variâncias v_{ij} pelas suas estimativas $\widehat{v}_{ij}$, representando por $\boldsymbol{V}$ a matriz com elementos principais $\widehat{v}_{ij}$, ordenados por ordem crescente dos respectivos índices h. Representemos por $\boldsymbol{Y}^w$ o vector de componentes $\widehat{y}_{ij}$, $(i, j) \in C$, também ordenados pelos respectivos índices h. Consegue-se assim agrupar os resultados de acordo com os níveis de exposição seguindo-se, dentro de cada nível, a ordem inicial das sub-populações.

Para

$$\boldsymbol{\eta}^w = (y_{\bullet\bullet}, \boldsymbol{\alpha}'^{(n-1)}, \boldsymbol{\gamma}'^{(s-1)})$$

ter-se-á o estimador (de mínimos quadrados pesados)

$$\widetilde{\boldsymbol{\eta}}^w = (\boldsymbol{X}_r^T \boldsymbol{V}^{-1} \boldsymbol{X}_r)^{-1} \boldsymbol{X}_r^T \boldsymbol{V}^{-1} \boldsymbol{Y}^d$$

onde $d = \#(C)$.

Se admitirmos que

$$\boldsymbol{Y}^d \sim N(\boldsymbol{\mu}^d, \boldsymbol{V})$$

ter-se-á

$$\widetilde{\boldsymbol{\eta}}^w \sim N(\boldsymbol{\eta}^w, (\boldsymbol{X}_r^T \boldsymbol{V}^{-1} \boldsymbol{X}_r)^{-1}).$$

É importante salientar que, nesta modelação, a matriz

$$\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\eta}}^w) = (\boldsymbol{X}_r^T \boldsymbol{V}^{-1} \boldsymbol{X}_r)^{-1}$$

é supostamente conhecida, o que permite utilizar os resultados sobre "*Variance Free Models, VFM*", obtidos por Mexia (1990).

Em particular com,

$$\boldsymbol{\Psi}^u = \boldsymbol{A}\boldsymbol{\eta}^w$$

e

$$\boldsymbol{W} = \boldsymbol{A}(\boldsymbol{X}_r^T \boldsymbol{V}^{-1} \boldsymbol{X}_r)^{-1} \boldsymbol{A}^T$$

ter-se-á

$$\widetilde{\boldsymbol{\Psi}}^u = \boldsymbol{A}\widetilde{\boldsymbol{\eta}}^w \sim N(\Psi, \boldsymbol{W})$$

vindo

$$Q(\boldsymbol{b}^u) = (\widetilde{\boldsymbol{\Psi}}^u - \boldsymbol{b}^u)^T \boldsymbol{W}^\dagger (\widetilde{\boldsymbol{\Psi}}^u - \boldsymbol{b}^u) \sim \chi^2_{r,\delta}$$

com $\mathrm{car}(\boldsymbol{W}) = r$ e $\boldsymbol{W}^\dagger$ a inversa generalizada de Moore-Penrose de $\boldsymbol{W}$.

Considere-se

$$\delta = (\boldsymbol{\Psi}^u - \boldsymbol{b}^u)^T \boldsymbol{W}^\dagger (\boldsymbol{\Psi}^u - \boldsymbol{b}^u).$$

Pode-se utilizar $Q(\boldsymbol{b}^u)$ para testar

$$H_0(\boldsymbol{b}^u) : \boldsymbol{\Psi}^u = \boldsymbol{b}^u.$$

Este teste é não distorcido e goza de dualidade, já que $H_0(\boldsymbol{b}^u)$ não é rejeitada pelo teste de nível q se e só se $\boldsymbol{b}^u$ pertence ao elipsóide de confiança de nível $1-q$ para $\boldsymbol{\Psi}^u$, dado por

$$(\boldsymbol{\Psi}^u - \widetilde{\boldsymbol{\Psi}}^u)^T \boldsymbol{W}^\dagger (\boldsymbol{\Psi}^u - \widetilde{\boldsymbol{\Psi}}^u) \leq x_{r,1-q}$$

onde $x_{r,1-q}$ é o quantil para a probabilidade $(1-q)$ da distribuição qui-quadrado central com r graus de liberdade.

Este teste pode ainda ser considerado como um teste $\mathcal{F}$ com uma infinidade de graus de liberdade para o erro. Assim, será **UMP** (Uniformemente Mais Potente) na família dos testes para $H_0(\boldsymbol{b}^u)$ cuja potência é função do parâmetro de não centralidade.

Consideremos $r = u$, $\boldsymbol{W}$ igual à sua inversa $\boldsymbol{W}^{-1}$ e definida positiva. Assim, $\delta = 0$ se e só se $H_0(\boldsymbol{b}^u)$ se verificar e o teste é estritamente não distorcido, ver Mexia (1992).

São particularmente importantes as hipóteses

$$\begin{cases} H_0(\boldsymbol{\alpha}'^{n-1}) : \boldsymbol{\alpha}'^{n-1} = \mathbf{0}^{n-1} \\ H_0(\boldsymbol{\gamma}'^{s-1}) : \boldsymbol{\gamma}'^{s-1} = \mathbf{0}^{s-1}. \end{cases}$$

Para as testar basta tomar como matrizes $\boldsymbol{A}$ as sub-matrizes de $\boldsymbol{I}_w$ formadas pelas linhas de 2 a n e de $n+1$ a w e aplicar a técnica anterior com $\boldsymbol{b}^{n-1} = \mathbf{0}^{n-1}$ e $\boldsymbol{b}^{s-1} = \mathbf{0}^{s-1}$. Como estas matrizes têm características $n-1$ e $s-1$,respectivamente, as matrizes $\boldsymbol{W}$ correspondentes também terão características $n-1$ e $s-1$, sendo definidas positivas, pelo que os testes serão estritamente não distorcidos.

Observe-se que, no caso geral, ver Mexia (1990), os intervalos de confiança simultâneos são dados por

$$pr[\bigcap_{\boldsymbol{d}^u}(|\boldsymbol{d}^{uT}\boldsymbol{\Psi}^u - \boldsymbol{d}^{uT}\widetilde{\boldsymbol{\Psi}}^u| \leq \sqrt{x_{r,1-q}\boldsymbol{d}^{uT}\boldsymbol{W}\boldsymbol{d}^u})] = 1 - q$$

os quais são fáceis de adaptar ao caso particular dos vectores $\boldsymbol{\alpha}'^{n-1}$ e $\boldsymbol{\gamma}'^{s-1}$.

Como

$$\alpha_n = -\sum_{i=1}^{n-1} \alpha_i$$

e

$$\gamma_s = -\sum_{j=1}^{s-1} \gamma_j,$$

tem-se que

$$\sum_{i=1}^{n} c_i\alpha_i = \sum_{i=1}^{n-1}(c_i - c_n)\alpha_i$$

e

$$\sum_{j=1}^{s} c'_j\gamma_j = \sum_{j=1}^{s-1}(c'_j - c'_s)\gamma_j,$$

isto é, as combinações lineares dos $\alpha_1, \ldots, \alpha_{n-1}$, respectivamente $\gamma_1, \ldots, \gamma_{s-1}$, incluem as combinações lineares dos $\alpha_1, \ldots, \alpha_n$, respectivamente $\gamma_1, \ldots, \gamma_s$.

Por outro lado, se

$$|\theta - \boldsymbol{d}^{uT}\widetilde{\boldsymbol{\Psi}}^u| > \sqrt{x_{r,1-q}\boldsymbol{d}^{uT}\boldsymbol{W}\boldsymbol{d}^u}$$

ter-se-á

$$pr(\boldsymbol{d}^{uT}\boldsymbol{\Psi}^u = \theta) < q$$

dizendo-se que $\boldsymbol{d}^{uT}\boldsymbol{\Psi}^u$ é significativamente diferente de θ ao nível q. Esta noção aplica-se directamente às combinações lineares dos $\alpha_1, \ldots, \alpha_n$, respectivamente $\gamma_1, \ldots, \gamma_s$, sendo em particular interessante ver quais os pares de efeitos $(\alpha_1, \ldots, \alpha_n)$, respectivamente $(\gamma_1, \ldots, \gamma_s)$, que diferem significativamente já que, $\alpha_i - \alpha_i' \neq 0$, respectivamente $\gamma_j - \gamma_j' \neq 0$, equivale a $\alpha_i \neq \alpha_i'$, respectivamente $\gamma_i \neq \gamma_i'$.

3.6.3 Estimadores de Máxima Verosimilhança

Vamos agora aplicar o algoritmo Zig-zag para obter estimadores de máxima verosimilhança para o modelo logit.

Tomemos

$$y_{ij} = logit(\widetilde{p}_{ij}) = ln\frac{\widetilde{p}_{ij}}{1-\widetilde{p}_{ij}}$$

onde $\widetilde{p}_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, s$, são probabilidades de ocorrência de acontecimentos dados. Estaremos pois perante um modelo com dois factores cruzados. No primeiro dos exemplos que adiante consideraremos, os acontecimentos correspondem à incidência (ocorrência) de Tuberculose e os níveis dos factores a países e anos. Quanto ao segundo exemplo, os acontecimentos correspondem à incidência de Sida e os níveis dos factores a distritos/regiões autónomas e anos.

É importante recordar que, neste caso,

$$Var(y_{ij}) \approx \frac{1}{N_{ij} \times p_{ij}}$$

com N_{ij} a dimensão da amostra.

Nos exemplos que consideraremos a amostra é constituída pela totalidade da população.

Nesta secção seguimos os trabalhos de Nunes *et al.* (2004A e 2004B), apresentando novamente o algoritmo Zig-zag para facilitar a compreensão dos exemplos que a seguir se discutem.

Vamos representar os efeitos correspondentes a países (no primeiro exemplo) e a distritos/regiões autónomas (no segundo exemplo) por f_i, $i = 1, \ldots, m$, e por g_j, $j = 1, \ldots, s$, os correspondentes a anos, reescrevendo o modelo na forma

$$y_{ij} = logit(\widetilde{p}_{ij}) = \beta_0 + \beta_1(f_i + g_j),$$

com $i = 1, \ldots, m$ e $j = 1, \ldots, s$.

Tomemos como valor inicial[5] para x_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, s$,

$$x_{ij}(0) = y_{i\bullet} + y_{\bullet j} - y_{\bullet\bullet}$$

[5]É importante referir que esta escolha resulta da prática que se tem tido na aplicação do algoritmo Zig-zag, ver Mexia *et al.* (1999 e 2001).

onde

$$\begin{cases} y_{i\bullet} = \frac{1}{s}\sum_{j=1}^{s} y_{ij} \\ y_{\bullet j} = \frac{1}{m}\sum_{i=1}^{m} y_{ij} \\ y_{\bullet\bullet} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{s} y_{ij}. \end{cases}$$

Considerando-se[6] $q_{ij} = \frac{1}{\mathrm{Var}(y_{ij})}$, a soma de quadrados dos resíduos toma a forma

$$S = \sum_{i=1}^{m}\sum_{j=1}^{s} q_{ij}(y_{ij} - \beta_0 - \beta_1 x_{ij}(0))^2 \quad = \sum_{i=1}^{m}\sum_{j=1}^{s} q_{ij}(y_{ij} - \beta_0 - \beta_1(f_i(0) + g_j(0)))^2.$$

Vamos agora, de forma muito sucinta, descrever os passos do algoritmo Zig-zag, apresentado na primeira secção deste capítulo, adaptado ao nosso caso particular.

O algoritmo Zig-zag é um algoritmo iterativo que, como o próprio nome indica, com minimizações alternadas, permite decompor o problema dentro de cada iteração $\imath$, $\imath = 1, 2, \ldots$, em dois problemas de mínimos quadrados:

- um problema de mínimos quadrados relativamente ao par de parâmetros (α, β), isto é, um problema em que se pretende minimizar a função objectivo em ordem a α e a β admitindo que se conhecem os valores x_{ij};

[6] O uso dos pesos inversamente proporcionais às variâncias é uma extensão do que se pode fazer nos mínimos quadrados generalizados. Assim, se se tiver um vector $\boldsymbol{Y}$ de valores da variável controlada com vector médio $\boldsymbol{X\beta}$ e matriz de covariância $\sigma^2\boldsymbol{C}$, com $\boldsymbol{C}$ conhecida e regular, tomando-se

$$\boldsymbol{Y'} = \boldsymbol{C}^{-1/2}\boldsymbol{Y}$$

e

$$\boldsymbol{X'} = \boldsymbol{C}^{-1/2}\boldsymbol{X}$$

é-se levado a minimizar

$$\|\boldsymbol{Y'} - \boldsymbol{X'\beta}\|^2 = (\boldsymbol{Y'} - \boldsymbol{X'\beta})^T(\boldsymbol{Y'} - \boldsymbol{X'\beta}) = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{C}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta}).$$

Caso $\boldsymbol{C}$ seja uma matriz diagonal, ter-se-á

$$\|\boldsymbol{Y'} - \boldsymbol{X'\beta}\|^2 = \sum_i \frac{1}{c_i}(y_i - \sum_j x_{ij}\beta_j)^2$$

que é uma expressão semelhante à da função S.

- um outro problema de mínimos quadrados relativamente ao par de vectores $(\boldsymbol{f}, \boldsymbol{g})$, com $\boldsymbol{f} = (f_1, \ldots, f_m)^T$ e $\boldsymbol{g} = (g_1, \ldots, g_s)^T$, isto é, em que se pretende minimizar a função objectivo admitindo que se conhece (α, β).

Como já foi referido, não conhecendo os valores das variáveis controladas, somos obrigados a obter estimativas iniciais, $x_{ij}(0)$, necessárias para iniciar a primeira iteração do algoritmo, que é definida pelos seguintes passos:

Passo 1 Consiste na minimização da função objectivo S em ordem aos parâmetros β_0 e β_1, usando as estimativas iniciais iniciais $x_{ij}(0)$.

Desta minimização resultam as estimativas:

$$\widetilde{\beta}_0(\imath) = y_\circ - \widetilde{\beta}_1(\imath) x_\circ(\imath) \qquad \text{e} \qquad \widetilde{\beta}_1(\imath) = \frac{s_{xy}(\imath)}{s_{xx}(\imath)}$$

onde

$$\begin{cases} y_\circ = \dfrac{\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{s} q_{ij} y_{ij}}{q^+} \\ \\ x_\circ(\imath) = \dfrac{\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{s} q_{ij} x_{ij}(0)}{q^+} \end{cases}$$

com

$$q^+ = \sum_{i=1}^{m}\sum_{j=1}^{s} q_{ij}$$

e

$$\begin{cases} s_{xx}(\imath) = \displaystyle\sum_{i=1}^{m}\sum_{j=1}^{s} q_{ij}(x_{ij}(0) - x_\circ(\imath))^2 \\ \\ s_{xy}(\imath) = \displaystyle\sum_{i=1}^{m}\sum_{j=1}^{s} q_{ij}(x_{ij}(0) - x_\circ(\imath))(y_{ij} - y_\circ). \end{cases}$$

Passo 2 Conhecendo as estimativas $\widetilde{\beta}_0(\imath)$ e $\widetilde{\beta}_1(\imath)$, obtidas no 1º passo do algoritmo, minimiza-se a função objectivo

$$S_{fg} = \sum_{i=1}^{m}\sum_{j=1}^{s} q_{ij}(y_{ij} - \widetilde{\beta}_0(\imath) - \widetilde{\beta}_1(\imath)(f_i + g_j))^2$$

em ordem aos vectores $\boldsymbol{f}$ e $\boldsymbol{g}$, obtendo-se o sistema

$$\left(\begin{array}{c|c} \boldsymbol{D}_1 & \boldsymbol{Q} \\ \hline \boldsymbol{Q}^T & \boldsymbol{D}_2 \end{array}\right)\left(\begin{array}{c} \boldsymbol{f} \\ \hline \boldsymbol{g} \end{array}\right) = \boldsymbol{v}(\imath)$$

onde

$$\boldsymbol{Q} = [q_{ij}]$$

é uma matriz do tipo $m \times s$,

$$\begin{cases} \boldsymbol{D}_1 = \operatorname{diag}(\sum_{j=1}^{s} q_{1j}, \ldots, \sum_{j=1}^{s} q_{mj}) \\ \\ \boldsymbol{D}_2 = \operatorname{diag}(\sum_{i=1}^{m} q_{i1}, \ldots, \sum_{i=1}^{m} q_{is}) \end{cases}$$

e o vector $\boldsymbol{v}(\imath) = (v_1(\imath), \ldots, v_{m+s}(\imath))^T$ tem por componentes

$$\begin{cases} v_i(\imath) = \frac{1}{\widetilde{\beta}_1(\imath)} \sum_{j=1}^{s} q_{ij}(y_{ij} - \widetilde{\beta}_0(\imath)) \ , \ i = 1, \ldots, m \\ \\ v_{m+j}(\imath) = \frac{1}{\widetilde{\beta}_1(\imath)} \sum_{i=1}^{m} q_{ij}(y_{ij} - \widetilde{\beta}_0(\imath)) \ , \ j = 1, \ldots, s. \end{cases}$$

A solução deste sistema produz as estimativas para os coeficientes f_i, $i = 1, \ldots, m$, e g_j, $j = 1, \ldots, s$, dadas por

$$\widetilde{f}_i(\imath) \ , \ i = 1, \ldots, m$$

e

$$\widetilde{g}_j(\imath) \ , \ j = 1, \ldots, s$$

pelo que

$$\widetilde{x}_{ij}(\imath) = \widetilde{f}_i(\imath) + \widetilde{g}_j(\imath).$$

Passo 3 Tendo-se obtido nos dois passos anteriores as estimativas $\widetilde{\beta}_0(\imath)$, $\widetilde{\beta}_1(\imath)$, $\widetilde{f}_i(\imath)$, $i = 1, \ldots, m$, e $\widetilde{g}_i(\imath)$, $j = 1, \ldots, s$, calculámos o valor estimado para a função objectivo

$$\widetilde{S}(\imath) = \sum_{i=1}^{m} \sum_{j=1}^{s} q_{ij}(y_{ij} - \widetilde{\beta}_0(\imath) - \widetilde{\beta}_1(\imath)(\widetilde{f}_i(\imath) + \widetilde{g}_j(\imath)))^2.$$

Passo 4 Este último passo do algoritmo consiste num processo de estandardização das estimativas de x_{ij}, que é efectuada com o objectivo de manter inalterados o mínimo e máximo de x_{ij}.

Considerando

$$\begin{cases} a = \min\{x_{ij}(0)\} \\ b = \max\{x_{ij}(0)\} \\ a(\imath) = \min\{\widetilde{x}_{ij}(\imath)\} \\ b(\imath) = \max\{\widetilde{x}_{ij}(\imath)\} \end{cases}$$

com $i = 1, \ldots, n$ e $j = 1, \ldots, s$, determina-se

$$\widetilde{w}_{ij}(\imath + 1) = \frac{b - a}{b(\imath) - a(\imath)}(\widetilde{x}_{ij}(\imath) - a(\imath)) + a.$$

Os valores obtidos nesta estandardização são utilizados no início da iteração seguinte, caso o valor da função objectivo, $\widetilde{S}(\imath)$, não tenha estabilizado, isto é, $x_{ij}(\imath) = \widetilde{w}_{ij}(\imath)$. Ver Mexia *et al.* (1999).

3.6.4 Exemplo 1 - Incidência de Tuberculose

A Tuberculose (TB) continua a ser uma preocupação mundial, sendo a causa de morte de cerca de 3 milhões de pessoas por ano.

A 24 de Março de 1882, Robert Koch descobriu o microrganismo responsável pela Tuberculose, que passaria a ser conhecido como o bacilo de Koch. Desde então o dia 24 de Março passou a ser assinalado como o Dia Mundial da Tuberculose. Este bacilo ataca principalmente os pulmões, no entanto a doença pode atingir vários órgãos do corpo humano: rins, coração, gânglios linfáticos, ossos, cérebro, etc.

A Organização Mundial de Saúde estima que um terço da população mundial, cerca de dois biliões de pessoas, esteja infectado com a bactéria da Tuberculose. Em cada ano há pelo menos oito milhões de novos casos, dos quais três milhões acabam por levar à morte.

A expansão do VIH e a crescente resistência aos medicamentos estão a agravar o impacto da doença, provocando o seu aumento em todo o mundo. Tal ocorre especialmente nos países mais pobres. No entanto, e depois de 40 anos de declínio, a Tuberculose começa a atacar os países industrializados. Na Europa, o caso que

estudaremos seguidamente, são os países de leste que apresentam a situação mais alarmante.

O estudo que em seguida apresentaremos é baseado na incidência da Tuberculose em 44 países Europeus, ou mais correctamente, da "WHO[7] European Region"de 1 de Janeiro de 1997 a 31 de Dezembro de 2000. Estes dados foram recolhidos nos seguintes relatórios: "Surveillance of Tuberculosis in Europe - Euro TB[8]"e "WHO Report - Global Tuberculosis Control".

As estimativas da população residente em cada um dos países em estudo foram obtidas na página WEB da "NIDI - The Netherlands Interdisciplinary Demographic Institute".

Foram feitas duas abordagens independentes à incidência da Tuberculose.

Começámos por um estudo global para os 44 países[9] no período de 4 anos. Posteriormente estratificámos o nosso universo, em primeiro lugar por sexo e seguidamente por idade. Nesta segunda estratificação, dividimos os dados por três grupos etários:

- Grupo Etário A (GE A) - indivíduos entre os zero e os 24 anos;
- Grupo Etário B (GE B) - indivíduos entre os 25 e os 54 anos;
- Grupo Etário C (GE C) - indivíduos com idade superior ou igual a 55 anos.

Temos então neste caso:

- $f_i \rightarrow$ o coeficiente i do factor de localização, país, $i = 1, \ldots, 44$;
- $g_j \rightarrow$ o coeficiente j do factor temporal, ano, $j = 1, \ldots, 4$.

Convém ainda salientar que, para esta primeira abordagem, os dados correspondentes à incidência da doença comportam todos os tipos de Tuberculose.

Numa segunda abordagem, apresentamos um estudo feito por sexo e idade (mantendo-se os mesmos grupos etários), com dados que reportam exclusivamente a incidência de Tuberculose Pulmonar. Apenas 40 países apresentaram este tipo de informação, pelo que não podemos fazer reais comparações entre as duas abordagens. Assim, neste caso, temos:

- $f_i \rightarrow$ o coeficiente i do factor de localização, país, $i = 1, \ldots, 40$;
- $g_j \rightarrow$ o coeficiente j do factor temporal, ano, $j = 1, \ldots, 4$.

[7]WHO é a abreviatura de "World Health Organization", designação da agência das Nações Unidas especializada em Saúde. Foi criada em 7 de Abril de 1948 e define como o seu principal objectivo "a obtenção por todas as pessoas do mais alto nível de saúde possível". A " WHO European Region"é constituída por 51 países europeus e asiáticos.

[8]Euro TB é a rede europeia para vigilância da Tuberculose. Foi criada em 1996 com o objectivo de melhorar o controlo da Tuberculose, sendo um centro colaborador da WHO.

[9]É importante salientar que, embora a "WHO European Region"fosse, neste período de tempo, constituída por 51 países, o facto de não existir informação disponível sobre todos os países para os casos que pretendíamos abranger, nomeadamente sobre sexo e idade, resolvemos trabalhar com os países para os quais toda a informação estivesse disponível.

Apresentação e Interpretação de Resultados

Em primeiro lugar é importante salientar que as estimativas que em seguida apresentamos foram obtidas na segunda iteração do algoritmo Zig-zag (uma vez que a partir desta o valor da função objectivo estabilizou); em segundo lugar, que estamos num caso homotópico, pelo que o relevante é a posição e não a escala.

Vamos começar por apresentar as estimativas obtidas utilizando simultaneamente um ambiente descritivo, com tabelas apresentando os diversos valores, e um ambiente gráfico[10], para uma mais fácil interpretação das situações estudadas.

Na interpretação destes dados utilizamos a técnica da Análise de Variância, ANOVA, estimando o erro a partir da soma de quadrados da interacção de mais alta ordem.

Completamos a nossa análise através da aplicação do método de comparação múltipla de Scheffé, ver Scheffé (1959).

Observe-se que este método é, ver Mexia (1987), robusto para a existência dum parâmetro de não centralidade positivo na soma de quadrados para o erro. No nosso caso, tal possibilidade decorre de se utilizar uma soma de quadrados de interacções para estimar o erro.

Aliás, esta propriedade de robustez estende-se, ver Mexia (1989), aos testes $\mathcal{F}$ utilizados em modelos de efeitos fixos.

1º Caso: Tuberculose - Todos os Tipos

Começamos por apresentar na Tabela 3.1 as estimativas para os parâmetros β_0 e β_1.

Tabela 3.1: **Estimativas para os parâmetros $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$**

Parâmetros	**Total**	**Sexo M**	**Sexo F**	**GE-A**	**GE-B**	**GE-C**
$\widetilde{\boldsymbol{\beta}}_{\mathbf{0}}$	0,03266	-0,01149	0,02515	-0,18505	-0,09239	0,02180
$\widetilde{\boldsymbol{\beta}}_{\mathbf{1}}$	1,00327	0,99779	1,00221	0,97526	0,98612	1,00203

Observe-se que, em todos estes casos, o valor do parâmetro $\widetilde{\boldsymbol{\beta}}_{\mathbf{1}}$ é aproximadamente igual a 1.

[10] O leitor constatará que o tamanho dos gráficos apresentados neste capítulo é relativamente pequeno. Tal facto deve-se a questões de organização do texto. Por esta razão todas as figuras deste capítulo encontram-se, em tamanho "normal", no Apêndice C.

Segue-se a Tabela 3.2 com as estimativas do factor temporal e o respectivo gráfico (Figura 3.1).

Tabela 3.2: **Estimativas para o factor temporal**, g

Ano	Total	Sexo M	Sexo F	GE-A	GE-B	GE-C
1997	-9,05	-8,78	-7,26	-8,95	-9,59	-9,02
1998	-9,02	-8,75	-7,24	-8,89	-9,57	-9,02
1999	-8,97	-8,76	-7,25	-8,90	-9,59	-9,04
2000	-8,93	-8,72	-7,21	-8,85	-9,55	-9,02

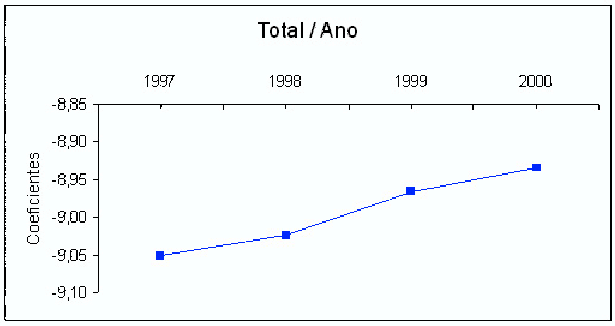

Figura 3.1: Estimativas para o factor temporal - Total.

Em termos globais, facilmente se conclui (ver Figura 3.1) que há um acréscimo, embora ligeiro, ao longo dos quatro anos em estudo.

Na Figura 3.2 observe-se o que se passa ao nível do Sexo.

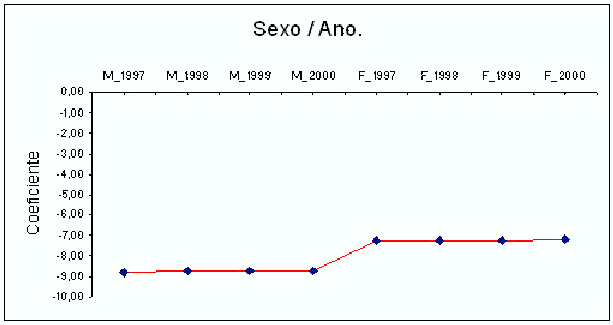

Figura 3.2: Estimativas para o factor temporal - Sexo.

É óbvia a diferença entre os valores para os dois sexos, registando-se ainda que em cada um deles praticamente não existe diferença ao longo dos quatro anos.

Quanto aos resultados por Grupo Etário, como podemos ver na Figura (3.3), existe uma diferença muito significativa entre os valores obtidos para o grupo etário B relativamente aos valores para os grupos A e C. Dentro de cada grupo não se verificam diferenças significativas ao longo do período em estudo.

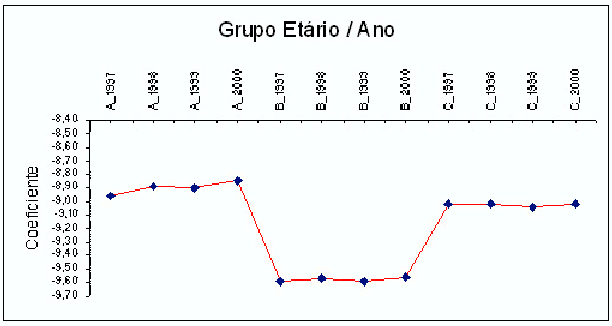

Figura 3.3: Estimativas para o factor temporal - Grupo Etário.

Na Tabela 3.3 encontramos as estimativas obtidas para o factor de localização.

Comecemos por analisar as estimativas do factor de localização em termos globais. Como veremos através dos próximos gráficos, os resultados mostraram uma clara divisão da Europa (WHO European Region) em três regiões:

- **Europa Ocidental**, onde a maioria dos países apresenta estimativas inferiores a $0,5$, com algumas excepções, das quais, infelizmente, Portugal faz parte;
- **Europa Central**, onde a maioria dos países apresenta estimativas que se situam entre $0,5$ e $1,5$;
- **Europa de Leste**, onde a maioria dos países apresenta estimativas superiores a $1,5$.

Para uma mais clara apresentação atribuímos cores a grupos de valores das estimativas:

- a cor verde corresponde às estimativas inferiores ou iguais a zero;
- a cor azul corresponde às estimativas com valores superiores a 0 e inferiores ou iguais a 1;
- a cor rosa corresponde às estimativas com valores entre 1 e 2;
- a cor vermelho corresponde às estimativas superiores ou iguais a 2.

Vejamos então os gráficos que ilustram as afirmações anteriores (Figura 3.4).

Tabela 3.3: **Estimativas para o factor de localização, f**

País	Total	Sexo M	Sexo F	GE-A	GE-B	GE-C
Albania	0,56	0,46	-1,36	-0,53	1,40	1,50
Alemanha	0,00	0,00	-2.07	-0,88	0,57	0,40
Arménia	1,21	1,48	-1,60	1,08	1,91	1,03
Áustria	0,24	0,24	-1,82	-0,81	0,89	0,67
Bélgica	0,00	0,06	-2,17	-0,90	0,62	0,42
Bósnia-Herzegovina	1,84	1,76	-0,10	0,64	2,27	2,72
Bulgária	1,25	1,35	-1,07	1,14	1,82	1,25
Cazaquistão	2,48	2,25	0,37	1,84	3,27	2,25
Croácia	1,26	1,28	-0,85	0,26	1,94	1,71
Dinamarca	-0,19	-0,31	-2,09	-0,60	0,65	-0,50
Eslováquia	0,61	0,59	-1,42	-1,34	1,14	1,61
Eslovénia	0,57	0,55	-1,45	-0,87	1,26	1,13
Espanha	0,56	0,50	-1,80	-0,05	1,13	0,48
Estónia	1,49	1,68	-0,87	0,25	2,51	1,56
Finlândia	-0,10	-0,24	-1,98	-2,34	-0,13	0,97
França	-0,09	-0,12	-2,10	-1,07	0,59	0,35
Geórgia	2,36	2,46	0,16	2,03	3,18	2,31
Grécia	-0,38	-0,33	-2,54	-1,17	0,00	0,00
Holanda	-0,31	-0,38	-2,27	-0,69	0,38	-0,28
Hungria	1,14	1,27	-1,14	-1,07	1,98	1,66
Irlanda	-0,08	-0,14	-2,06	-1,03	0,45	0,66
Islândia	-0,95	-1,59	-2,39	-1,15	-0,44	-0,01
Israel	-0,26	-0,33	-2,24	-1,49	0,40	0,71
Itália	-0,40	-0,42	-2,46	-1,43	0,11	0,03
Letónia	1,91	2,09	-0,52	1,16	2,84	1,80
Lituânia	1,90	2,00	-0,30	0,89	2,73	2,25
Luxemburgo	-0,23	-0,12	-2,40	-0,90	0,43	0,09
Macedónia	0,94	0,88	-1,03	0,52	1,57	1,36
Malta	-1,03	-0,81	-3,63	-1,44	-1,12	0,01
Moldávia (Rep.)	1,69	1,82	-0,67	0,92	2,68	1,58
Noruega	-0,83	-1,01	-2,65	-1,46	-0,27	-0,40
Polónia	0,98	1,05	-1,22	-0,86	1,81	1,62
Portugal	1,39	1,49	-0,82	0,58	2,29	1,46
Reino Unido	-0,13	-0,25	-2,02	-0,75	0,52	0,15
República Checa	0,31	0,27	-1,72	-1,72	0,81	1,14
Roménia	2,22	2,35	-0,06	1,33	2,94	1,46
Rússia (Fed.)	1,96	2,16	-0,64	1,33	2,96	1,46
Suécia	-0,86	-1,10	-2,63	-1,57	-0,32	-0,47
Suíça	-0,20	-0,28	-2,15	-0,79	0,40	0,07
Tajiquistão	1,21	1,08	-0,66	0,65	2,33	1,30
Turquemenistão	1,91	1,88	-0,10	1,44	2,95	1,89
Ucrânia	1,60	1,72	-0,84	0,72	2,51	1,43
Usbequistão	1,61	1,50	-0,37	0,90	2,63	1,83
Jugoslávia	0,89	0,86	-1,13	-0,43	1,57	1,46

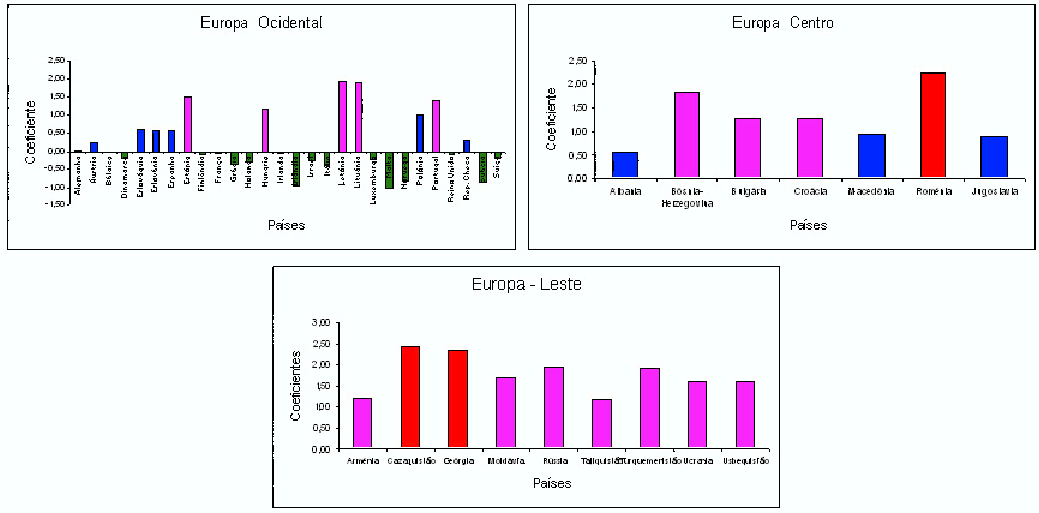

Figura 3.4: Estimativas para o factor de localização - Total - WHO European Region.

Considerando, do grupo que chamamos "Europa Ocidental", apenas os países que fazem parte da Comunidade Europeia, observamos que as diferenças ainda são maiores.

A maioria dos países que formavam a Comunidade Europeia no período em estudo (1997-2000) apresentam estimativas inferiores a zero, com apenas três excepções, a Áustria que apresenta um valor inferior a $0,5$, a Espanha com um valor pouco acima de $0,5$ e Portugal que, infelizmente, apresenta um valor muito próximo de $1,5$, estando deste modo muito afastado da média dos países da Comunidade Europeia.

Se a este grupo juntarmos todos os países que constituem a Comunidade Europeia em 2005, a situação altera-se ligeiramente, continuando Portugal a ser um dos países com pior resultado.

Observemos graficamente estas duas situações na Figura (3.5).

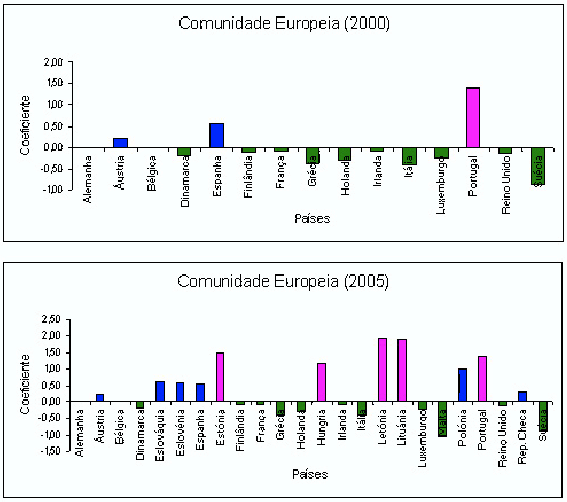

Figura 3.5: Estimativas para o factor de localização - Total - CE (2000 e 2005).

Vamos agora observar, na Figura 3.6, o que se passa quando se estuda a incidência da Tuberculose estratificando por sexo.

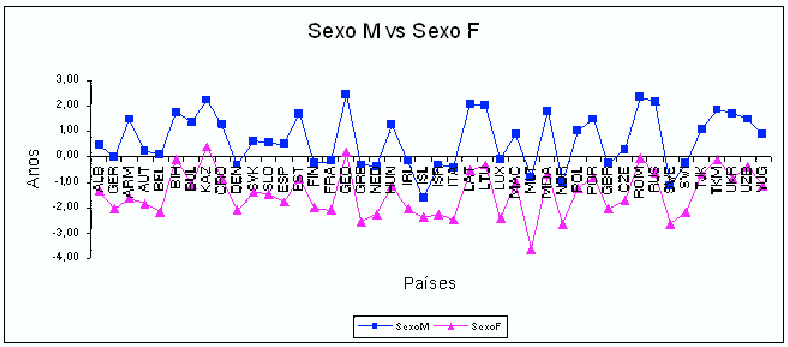

Figura 3.6: Estimativas para o factor de localização - Sexo.

A diferença entre os valores obtidos para os dois sexos é indiscutível. Não ficam dúvidas que o sexo masculino apresenta valores muito mais preocupantes.

No que respeita ao estudo das incidências por grupo etário, embora as diferenças não sejam tão significativas e tão generalizadas a todos os países, podemos concluir (ver Figura 3.7) que na maioria dos países o grupo etário A apresenta valores inferiores aos outros dois grupos.

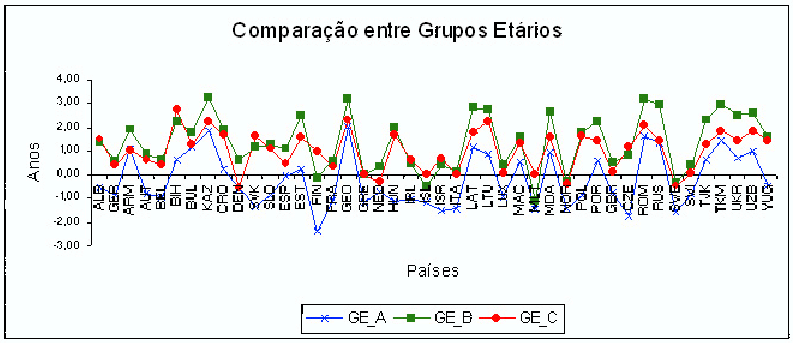

Figura 3.7: Estimativas para o factor de localização - Grupo Etário.

Quanto ao ajustamento do modelo podemos comprovar, observando na Tabela 3.4 os valores do coeficiente de determinação, que o método dos mínimos quadrados estruturados permitiu ajustar com precisão um modelo logístico a partir de dados apenas sobre a incidência da doença.

Tabela 3.4: **Coeficiente de Determinação, R^2**

	Total	**Sexo M**	**Sexo F**	**GE A**	**GE B**	**GE C**
R^2	0,987	0,990	0,975	0,995	0,991	0,979

Resultados da Análise de Variância

Comecemos por apresentar a tabela ANOVA (Tabela 3.5) para o caso Sexo vs Ano.

Tabela 3.5: **ANOVA - Sexo vs Ano**

OV	SQ	GL	QM	$\mathcal{F}$
Ano	0,00403	3	0,00134	57,47
Sexo	4,57493	1	4,57493	195517,9
Sexo×*Ano*	0,00007	3	2,3399E-05	-

O teste de correlação múltipla de Scheffé conduziu à seguinte diferença significativa:

$$ds_1(ano) = 0,05246$$

donde se obtém a seguinte relação entre os quatro anos em estudo

$$1^o\text{ano},\ 2^o\text{ano},\ 3^o\text{ano} \gg 4^o\ \text{ano}.$$

Tabela 3.6: **ANOVA - Grupo Etário vs Ano**

OV	SQ	GL	QM	$\mathcal{F}$
Ano	0,003692	3	0,001231	2,505131
Idade	1,032241	2	0,516121	1050,624
Idade×*Ano*	0,002948	6	0,000491	-

O segundo caso analisado foi Grupo Etário vs Ano, cuja tabela ANOVA é apresentada na Tabela 3.6.

Neste caso a diferença significativa obtida através do teste de Scheffé é

$$ds_2(idade) = 0,200999$$

pelo que a relação entre os três grupos etários é a seguinte

$$\text{Grupo Etário B} \ll \text{Grupo Etário C} \ll \text{Grupo Etário A}.$$

Na Tabela 3.7 apresentamos os resultados da análise para o caso Sexo vs País.

Tabela 3.7: **ANOVA - Sexo vs Países**

OV	SQ	GL	QM	$\mathcal{F}$
Países	80,40652	43	1,869919	24,42987
Sexo	97,28391	1	97,28391	1270,982
Sexo×*Países*	3,29132	43	0,076542	-

O teste de Scheffé permitiu obter a seguinte diferença significativa:

$$ds_3(países) = 4,702943$$

.

Neste caso existem inúmeras relações entre os países em estudo, indicamos apenas as mais significativas:

- Cazaquistão $\gg$ Bélgica, Dinamarca, Finlândia, França, Grécia, Holanda, Irlanda, Islândia, Israel, Itália, Luxemburgo, Malta, Noruega, Reino Unido, Suécia, Suíça;
- Geórgia $\gg$ Dinamarca, Finlândia, França, Grécia, Holanda, Irlanda, Islândia, Israel, Itália, Luxemburgo, Malta, Noruega, Reino Unido, Suécia, Suíça;
- Islândia $\ll$ Bósnia-Herzegovina, Cazaquistão, Estónia, Geórgia, Letónia, Lituânia, Moldávia, Roménia, Rússia, Turquemenistão, Ucrânia, Usbequistão;
- Malta $\ll$ Bósnia-Herzegovina, Bulgária, Cazaquistão, Croácia, Estónia, Geórgia, Letónia, Lituânia, Moldávia, Portugal, Roménia, Rússia, Tajiquistão, Turquemenistão, Ucrânia, Usbequistão.

Tabela 3.8: **ANOVA - Grupo Etário vs Países**

OV	SQ	GL	QM	$\mathcal{F}$
Países	124,85244	43	2,903545	13,22353
Idade	60,75254	2	30,37627	138,3418
Idade$\times Países$	18,88337	86	0,219574	-

Temos finalmente o caso Grupo Etário vs País, cujos resultados são apresentados na Tabela 3.8.

Neste último caso testámos não só os países mas também a idade (grupo etário). Os resultados foram os seguintes:

$$ds_4(idade) = 10,98055$$

e

$$ds_5(países) = 9,370575.$$

Relativamente à idade, este teste conduziu à seguinte relação entre os três grupos etários:

$$\text{Grupo Etário A} \ll \text{Grupo Etário C} \ll \text{Grupo Etário B}.$$

No que respeita aos países, tal como aconteceu no caso anterior, apresentamos apenas as relações mais significativas:

- Cazaquistão $\gg$ Malta, Noruega, Suécia;
- Geórgia $\gg$ Malta, Noruega, Suécia;
- Malta $\ll$ Cazaquistão, Geórgia, Roménia.

2º Caso: Tuberculose Pulmonar

Também neste caso as estimativas apresentadas foram obtidas no fim da segunda iteração do algoritmo.

Tal como no caso anterior, começamos pelas estimativas dos parâmetros β_0 e β_1 (ver Tabela 3.9).

Tabela 3.9: **Estimativas para os parâmetros $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$**

Parâmetros	**AM**	**AF**	**BM**	**BF**	**CM**	**CF**
$\widetilde{\boldsymbol{\beta_0}}$	-2,23412	-1,84962	-2,20310	-2,24199	-2,02551	-2,36418
$\widetilde{\boldsymbol{\beta_1}}$	0,74015	0,78752	0,69612	0,73343	0,7441	0,74191

Neste caso o valor do parâmetro $\widetilde{\boldsymbol{\beta_1}}$ é aproximadamente igual a $0,7$, ou seja a taxa de variação é inferior.

Na Tabela 3.10 podemos encontrar as estimativas para o factor temporal.

Tabela 3.10: **Estimativas para o factor temporal**, g

Ano	GEAM	GEAF	GEBM	GEBF	GECM	GECF
1997	-12,09	-12,22	-10,63	-11,57	-9,83	-10,86
1998	-12,05	-12,09	-10,67	-11,61	-9,88	-10,88
1999	-11,80	-11,93	-10,42	-11,41	-9,86	-10,83
2000	-12,05	-11,93	-10,81	-11,56	-10,09	-10,94

Observemos graficamente (ver Figura 3.8), por grupo etário, os valores obtidos para ambos os sexos.

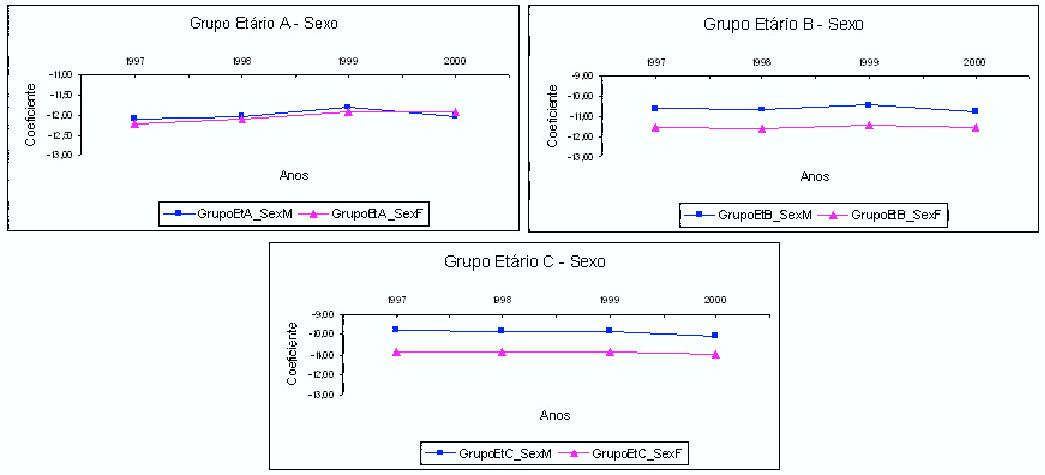

Figura 3.8: Estimativas para o factor temporal - Grupos Etários / Sexo.

No caso do grupo etário A as duas linhas sobrepõem-se praticamente ao longo dos quatro anos. Durante este período, verificamos um crescimento nos primeiros três anos observando-se um ligeiro decréscimo no último.

Nos grupos etários B e C é notória a diferença entre as duas linhas que representam os dois sexos, e o sexo feminino apresenta melhor valor.

As estimativas para o factor de localização encontram-se na Tabela 3.11.

Tabela 3.11: **Estimativas para o factor de localização, f**

Ano	AM	AF	BM	BF	CM	CF
Albania	0,69	0,78	1,05	1,26	1,26	1,68
Alemanha	0,00	0,00	0,00	0,00	0,00	0,00
Arménia	3,54	1,36	2,45	1,49	1,26	0,57
Áustria	1,41	1,58	1,47	1,67	1,27	1,73
Azerbaijão	1,46	0,18	1,95	2,21	0,06	0,76
Bélgica	-0,05	-0,04	0,07	0,35	0,26	-0,21
Bósnia-Herzegovina	2,12	2,40	2,19	2,38	2,22	3,46
Cazaquistão	3,69	3,69	3,71	4,18	2,82	2,72
Croácia	2,24	2,23	2,56	2,42	2,00	2,61
Dinamarca	0,00	0,40	-0,56	0,23	-1,36	-0,59
Eslováquia	-1,44	-0,60	0,88	-0,12	0,68	1,55
Eslovénia	-0,29	0,87	1,35	1,12	0,81	1,30
Estónia	1,52	1,38	3,12	2,46	1,95	1,07
Finlândia	-1,82	-1,86	-0,19	-0,70	0,76	1,06
França	0,03	0,24	0,02	0,23	-0,14	0,31
Geórgia	2,47	1,64	1,94	1,85	0,81	0,12
Grécia	-0,33	0,06	-0,66	-0,83	-0,67	-0,68
Holanda	0,05	-0,07	-1,45	-0,61	-1,78	-1,82
Hungria	-0,95	-0,43	1,51	0,64	0,49	0,17
Irlanda	-0,18	0,12	-0,33	0,40	-0,08	0,68
Israel	-0,26	-0,22	0,03	0,18	0,41	0,92
Itália	-0,15	-0,16	-0,60	-0,25	-0,34	-0,15
Letónia	2,91	2,25	3,37	2,83	2,13	1,18
Lituânia	2,10	1,93	3,39	2,87	2,51	2,10
Luxemburgo	1,59	1,90	1,17	0,88	0,12	1,35
Macedónia	1,04	1,58	1,60	1,73	1,14	1,15
Malta	-0,01	0,33	-0,71	-0,81	0,55	0,52
Moldávia (Rep.)	2,19	1,29	1,94	1,81	0,66	0,15
Noruega	-0,99	-0,65	-2,42	-1,06	-1,01	-0,43
Polónia	-0,12	0,26	1,88	1,32	1,29	1,59
Portugal	2,43	2,33	2,57	2,41	1,54	1,05
Quirguistão	2,97	2,61	3,46	3,52	2,36	2,45
Reino Unido	-0,57	-0,21	-1,46	-0,56	-1,19	-1,10
República Checa	-1,43	-0,78	0,75	-0,05	0,65	0,89
Roménia	3,72	3,56	4,18	3,68	2,75	2,31
Rússia (Fed.)	1,48	0,74	1,79	1,17	0,62	-0,25
Suécia	-1,09	-0,51	-2,51	-0,91	-1,90	-1,23
Suíça	-0,20	-0,05	-1,57	-0,58	-1,41	-1,46
Turquemenistão	2,30	1,78	3,00	3,35	2,05	2,72
Ucrânia	2,27	2,02	3,10	2,26	1,92	1,21

Na Figura 3.9 faz-se uma comparação dos valores obtidos em cada grupo etário para os dois sexos.

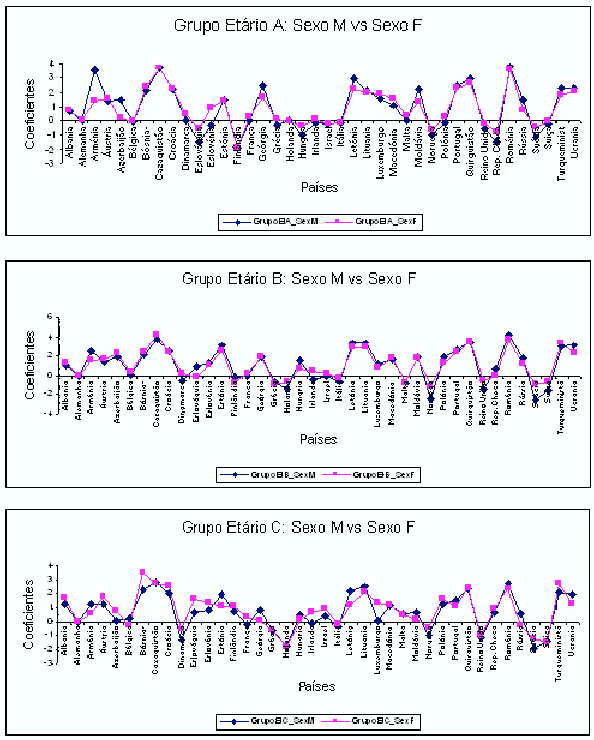

Figura 3.9: Estimativas para o factor de localização - Grupos Etários / Sexo.

Como podemos observar, no grupo etário A não se verificam diferenças muito significativas entre os valores obtidos para ambos os sexos. No entanto, o sexo masculino apresenta valores piores nos países em que a situação é mais grave, verificando-se precisamente o oposto nos países em que a Tuberculose apresenta uma menor incidência.

No grupo etário B o sexo feminino apresenta, na generalidade dos quarenta países, valores superiores aos do sexo masculino.

Finalmente, o grupo etário C não apresenta grande diferenças entre os valores obtidos para os dois sexos.

Tal como no caso anterior, obtivemos um ajustamento com precisão. Na Tabela 3.12 podemos observar os valores do coeficiente de determinação.

Tabela 3.12: **Coeficiente de Determinação**, $\boldsymbol{R^2}$

	AM	**AF**	**BM**	**BF**	**CM**	**CF**
$\boldsymbol{R^2}$	0,915	0,919	0,931	0,899	0,891	0,956

Resultados da Análise de Variância

À semelhança do que se fez para o caso anterior, apresentamos na Tabela 3.13 os resultados da análise de variância efectuada.

Tabela 3.13: **ANOVA - Grupo Etário+Sexo vs Países**

OV	SQ	GL	QM	$\mathcal{F}$
Idade	5,700643	2	2,850322	24,590907
Sexo	0,063302	1	0,063302	0,546133
País	396,0341	39	10,154719	87,608972
Idade $\times Sexo$	0,140091	2	0,070045	0,0604311
Idade $\times País$	60,49379	78	0,775561	6,691089
Sexo $\times País$	12,2147	39	0,313197	2,702084
Idade $\times Sexo \times País$	9,04	78	0,115910	-

Desta análise de variância concluímos que existe uma relação significativa entre grupo etário e país. O teste de Scheffé para cada um destes items conduziu às seguintes diferenças significativas:

$$ds_6(idade) = 7,618892$$

e

$$ds_7(país) = 6,483857$$

donde se retiram as seguintes relações significativas:

$$\text{Grupo Etário C} \ll \text{Grupo Etário A} \ll \text{Grupo Etário B}.$$

e

- Cazaquistão $\gg$ Dinamarca, Eslováquia, Eslovénia, Finlândia, França, Grécia, Holanda, Hungria, Irlanda, Israel, Itália, Luxemburgo, Malta, Noruega, Polónia, Reino Unido, Rep. Checa, Rússia, Suécia, Suíça;
- Bósnia-Herzegovina $\gg$ Dinamarca, Eslováquia, Finlândia, França, Grécia, Holanda, Hungria, Irlanda, Israel, Itália, Malta, Reino Unido, Rep. Checa, Rússia, Suécia, Suíça.

3.6.5 Exemplo 2 - Incidência de SIDA

O Síndroma da Imunodeficiência Adquirida (SIDA) foi identificado pela primeira vez em 1981 e o seu agente, Vírus da Imunodeficiência Humana (VIH), foi descoberto em 1984. A SIDA é um problema de saúde pública importante, estimando a Organização Mundial de Saúde que o número de indivíduos infectados no mundo em 2003 atingisse os 37,8 milhões, dos quais 35,7 milhões adultos e 2,1 milhões crianças com idade inferior a 15 anos.

O estudo apresentado baseia-se nos dados de incidência de SIDA em Portugal de 1 de Janeiro de 1990 até 31 de Dezembro de 2002, casos diagnosticados, fornecidos

pelo Centro de Vigilância Epidemiológica das Doenças Transmissíveis do Instituto Nacional de Saúde. As estimativas da população residente em Portugal foram obtidas a partir das publicações do Gabinete de Estudos Demográficos do Instituto Nacional de Estatística entre 1990 e 2002.

Temos então:

- $f_i \rightarrow$ o coeficiente i do factor de localização, distrito ou região autónoma, $i = 1, \ldots, 20$;

- $g_j \rightarrow$ o coeficiente j do factor temporal, ano, $j = 1, \ldots, 13$.

Vamos começar por apresentar os resultados da aplicação do modelo e o consequente ajustamento após quatro iterações do algoritmo Zig-zag.

Tabela 3.14: **Estimativas para os parâmetros $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$**

Parâmetros	**Estimativa**
$\widetilde{\boldsymbol{\beta_0}}$	0,10196
$\widetilde{\boldsymbol{\beta_1}}$	1,01075

Como podemos observar na Tabela 3.14, também neste caso a estimativa para a taxa de variação tem um valor aproximadamente igual a 1.

Na Tabela 3.15 e na Figura 3.10 encontramos as estimativas para o factor temporal.

Tabela 3.15: **Estimativas para o factor temporal - SIDA**

Ano	Estimativa
1990	-11,31
1991	-11,19
1992	-10,88
1993	-10,60
1994	-10,45
1995	-10,31
1996	-10,12
1997	-10,12
1998	-10,11
1999	-10,08
2000	-10,19
2001	-10,20
2002	-10,26

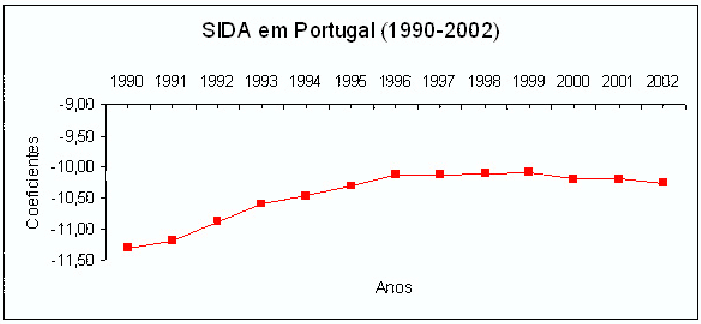

Figura 3.10: Estimativas para o factor temporal - SIDA.

As estimativas do factor de localização, os seus valores são apresentados na Tabela 3.16.

Tabela 3.16: **Estimativas para o factor de localização - SIDA**

Distrito/Região Autónoma	Estimativa
Açores	-0,32
Aveiro	-0,29
Beja	-0,32
Braga	-0,59
Bragança	-0,19
Castelo Branco	-0,21
Coimbra	-0,02
Évora	0,41
Faro	0,96
Guarda	-0,50
Leiria	0,28
Lisboa	1,69
Madeira	0,00
Portalegre	-0,65
Porto	1,30
Santarém	0,15
Setúbal	1,65
Viana do Castelo	-0,20
Vila Real	-0,63
Viseu	-0,35

A Figura (3.11) auxilia-nos na interpretação destes valores.

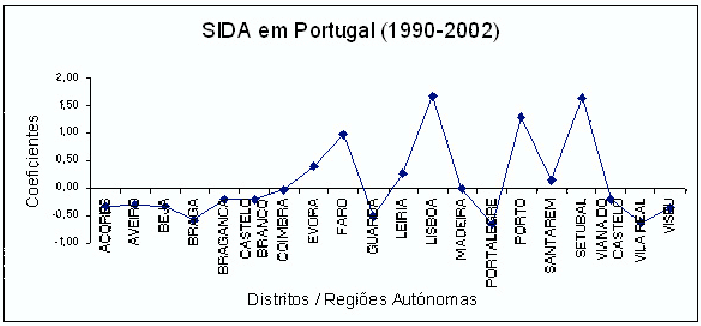

Figura 3.11: Estimativas para o factor de localização - SIDA.

O método dos mínimos quadrados estruturados permitiu, mais uma vez, ajustar com precisão um modelo logístico a partir de dados apenas sobre a incidência de uma doença. O valor do coeficiente de determinação obtido foi de $R^2 \approx 0,90$. Ultrapassou-se assim a dificuldade associado ao facto da SIDA não ser uma doença de notificação obrigatória no período em estudo.

Mais uma vez salientamos que estamos perante uma situação homotópica, pelo que a relevância reside na ordenação e nas diferenças significativas.

O modelo espacio-temporal ajustado permite concluir que:

- a SIDA é, em todo o país, um problema sério de saúde pública sendo os distritos de Lisboa, Setúbal e Porto (por esta ordem) os que apresentam valores de incidência estimados mais elevados;

- terá havido um ligeiro acréscimo da incidência da doença de 1990 a 1999, verificando-se ao longo dos restantes três anos um, também ligeiro, decréscimo. Esta conclusão confirma a obtida por Oliveira e Mexia (2004) e afasta os cenários epidémicos propostos por Carvalho e Diamantino (1993) e Amaral (2000).

Interessará regressar ao estudo deste problema quando a doença for de notificação obrigatória. Poder-se-á então acompanhar com mais precisão a evolução da situação. Nesse estudo poderá ser útil o modelo espacio-temporal pois assenta em poucos pressupostos, possuindo robustez; ao contrário do que se verifica em modelos que incorporam estimativas da duração do período de incubação, como é o caso do Back-Calculation, que são muito sensíveis à distribuição escolhida para essa duração (ver Carvalho e Diamantino, 1993).

Capítulo 4

Medições Aproximadas e Enviesamento

4.1 O Problema

Como já várias vezes o referimos, é notório que os modelos logit são largamente utilizados para exprimir a probabilidade de ocorrências duma determinada doença como função de efeitos ambientais.

Duma forma geral, as exposições são medidas em estações de medição ambientais previamente fixadas. Estas exposições são então consideradas como representativas dum determinado conjunto populacional residente nas proximidades de cada estação. Como consequência, é fundamental ter em consideração não só os erros de medição mas também o enviesamento (bias) existente nessas medições.

O objectivo fundamental deste capítulo consiste em apresentar um modelo que permita examinar o que pode acontecer quando os erros de medição não são desprezáveis.

Recorde-se que no modelo

$$y = logit(p) = \beta_0 + \beta_1 f$$

onde f representa a exposição, β_1 representa a taxa de variação de y com f. Este parâmetro tem, pois, uma importância central. Como veremos, o utilizar-se medidas aproximadas da exposição leva a um enviesamento por defeito na estimação de β_1, tendo-se

$$E(\widetilde{\beta}_1) < \beta_1.$$

Quando se utiliza o método dos mínimos quadrados estruturados, os estimadores das exposições podem ser considerados como medidas aproximadas das mesmas.

Este é o resultado fundamental deste capítulo, traduzindo a perda de sensibilidade na variação da exposição, a qual, por sua vez, resulta da falta de precisão nas medições efectuadas.

Depois de obtermos este resultado, utilizaremos a aproximação de Edgeworth para obter limites de confiança para o enviesamento.

4.2 Valores Médios para o Enviesamento

Vamos então assumir que, com a exposição f, a probabilidade de um indivíduo estar infectado é dada por

$$p = \frac{e^{\beta_0+\beta_1 f}}{1+e^{\beta_0+\beta_1 f}}.$$

Sejam $\widehat{\beta}_1$ e $\widetilde{\beta}_1$, respectivamente, os estimadores de máxima verosimilhança obtidos a partir das exposições exactas $f_1, \ldots, f_n$ e das exposições aproximadas

$$\tilde{f}_i = f_i + \varepsilon_i,\ i = 1, \ldots, n,$$

sendo $\varepsilon_1, \ldots, \varepsilon_n$ independentes e identicamente distribuídos, com valor médio nulo e variância σ^2_ε.

Suponhamos que, em cada grupo i, $i = 1, \ldots, n$, são observados n_i indivíduos, dos quais x_i estão infectados. Consideremos ainda que, para o grupo i, $\tilde{f}_i$ é a exposição média aproximada (medida) e f_i é a exposição média real. Nestas condições podemos considerar os seguintes modelos:

1. *modelo "aproximado"*: $logit(\frac{x_i}{n_i}) = \beta_0+\beta_1(\tilde{f}_i)$ - "aproximado"porque é baseado na exposição aproximada $\tilde{f}_i$;
2. *modelo "correcto"*: $logit(\frac{x_i}{n_i}) = \beta_0 + \beta_1(f_i)$ - "correcto"porque partimos do pressuposto de que todos os indivíduos de cada grupo i, $i = 1, \ldots, n$, têm igual risco de infecção.

Tomemos

$$\begin{cases} y_i = \text{logit}(\frac{x_i}{n_i}) \\ \\ v_i = \text{Var}(y_i) \end{cases}$$

bem como

$$\boldsymbol{X}^T \boldsymbol{D}^{-1} \boldsymbol{X} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

onde

$$a = \sum_{i=1}^{n} \frac{1}{v_i}, \qquad b = \sum_{i=1}^{n} \frac{f_i}{v_i}, \qquad c = \sum_{i=1}^{n} \frac{f_i^2}{v_i}$$

e

$$d = ac - b^2.$$

Analogamente, obtemos

$$\tilde{\boldsymbol{X}}^T \boldsymbol{D}^{-1} \tilde{\boldsymbol{X}} = \begin{pmatrix} a & \tilde{b} \\ \tilde{b} & \tilde{c} \end{pmatrix}$$

vindo

$$\tilde{b} = b + \sum_{i=1}^{n} \frac{\varepsilon_i}{v_i},$$

$$\tilde{c} = \sum_{i=1}^{n} \frac{(f_i + \varepsilon_i)^2}{v_i} = c + 2\sum_{i=1}^{n} \frac{f_i \varepsilon_i}{v_i} + \sum_{i=1}^{n} \frac{\varepsilon_i^2}{v_i}$$

e

$$\tilde{d} = a\tilde{c} - \tilde{b}^2.$$

Logo,

$$(\boldsymbol{X}^T \boldsymbol{D}^{-1} \boldsymbol{X})^{-1} = \frac{1}{d} \begin{pmatrix} c & -b \\ -b & a \end{pmatrix}$$

e, portanto,

$$(\tilde{\boldsymbol{X}}^T \boldsymbol{D}^{-1} \tilde{\boldsymbol{X}})^{-1} = \frac{1}{\tilde{d}} \begin{pmatrix} \tilde{c} & -\tilde{b} \\ -\tilde{b} & a \end{pmatrix}.$$

Por outro lado, temos

$$\boldsymbol{X}^T \boldsymbol{D}^{-1} \boldsymbol{y}^n = \begin{pmatrix} u \\ v \end{pmatrix}$$

onde

$$u = \sum_{i=1}^{n} \frac{y_i}{v_i} \qquad \text{e} \qquad v = \sum_{i=1}^{n} \frac{f_i y_i}{v_i}$$

bem como

$$\tilde{\boldsymbol{X}}^T \boldsymbol{D}^{-1} \boldsymbol{y}^n = \begin{pmatrix} u \\ \tilde{v} \end{pmatrix}$$

com

$$\tilde{v} = v + \sum_{i=1}^{n} \frac{\varepsilon_i y_i}{v_i}.$$

Assim, os estimadores serão

$$\widehat{\beta}_1 = \frac{av - bu}{d},$$

independente de $\boldsymbol{\varepsilon}^n$, e

$$\widetilde{\beta}_1 = \frac{a\tilde{v} - \tilde{b}u}{\tilde{d}}.$$

Note-se que dada uma função $g(y_1, y_2) = \frac{y_1}{y_2}$, tem-se:

$$\begin{cases} \text{g'}_{y_1}(y_1, y_2) = \frac{1}{y_2} \\ \\ \text{g'}_{y_2}(y_1, y_2) = -\frac{y_1}{y_2}. \end{cases}$$

Deste modo, podemos escrever

$$\widetilde{\beta}_1 \approx \widehat{\beta}_1 + \frac{1}{d}[a(\tilde{v} - v) - (\tilde{b} - b)u] - \frac{\widehat{\beta}_1}{d}(\tilde{d} - d).$$

Relembrando que ε_i é independente de y_i e que $\mathrm{E}[\varepsilon_i] = 0$, obtemos

$$\mathrm{E}(\tilde{v} - v) = \mathrm{E}(\sum_{i=1}^{n} \frac{\varepsilon_i y_i}{v_i}) = 0$$

e

$$\mathrm{E}(\tilde{b} - b) = \mathrm{E}(\sum_{i=1}^{n} \frac{\varepsilon_i}{v_i}) = 0.$$

Como $\widehat{\beta}_1$ e $(\tilde{d} - d)$ são independentes, temos que

$$\mathrm{E}(\widetilde{\beta}_1 - \widehat{\beta}_1) \approx -\frac{\beta_1}{d}\mathrm{E}(\tilde{d} - d).$$

Uma vez que

$$\tilde{d} - d = a(\tilde{c} - c) - (\tilde{b}^2 - b^2),$$

$$\mathrm{E}(\tilde{c} - c) = \sigma_\varepsilon^2 \sum_{i=1}^{n} \frac{1}{v_i},$$

$$\tilde{b}^2 - b^2 = 2b \sum_{i=1}^{n} \frac{\varepsilon_i}{v_i} + (\sum_{i=1}^{n} \frac{\varepsilon_i}{v_i})^2$$

e que

$$\mathrm{E}(\tilde{b}^2 - b^2) = \mathrm{Var}(\sum_{i=1}^{n} \frac{\varepsilon_i}{v_i}) = \sigma_\varepsilon^2 \sum_{i=1}^{n} \frac{1}{v_i^2}$$

obtemos

$$\mathrm{E}(\tilde{d} - d) = \sigma_\varepsilon^2[(\sum_{i=1}^{n} \frac{1}{v_i})^2 - \sum_{i=1}^{n} \frac{1}{v_i^2}]$$

e, finalmente, o nosso principal resultado:

$$\mathrm{E}(\widetilde{\beta}_1 - \widehat{\beta}_1) \approx -\beta_1 \frac{[(\sum_{i=1}^{n} \frac{1}{v_i})^2 - \sum_{i=1}^{n} \frac{1}{v_i^2}]}{(\sum_{i=1}^{n} \frac{1}{v_i})(\sum_{i=1}^{n} \frac{f_i^2}{v_i}) - (\sum_{i=1}^{n} \frac{f_i}{v_i})^2} \sigma_\varepsilon^2 < 0.$$

O que nos leva a concluir que o bias é sempre negativo.

Um resultado desta natureza expressa a perda de sensibilidade na variação da exposição, a qual resulta da falta de precisão nas medições efectuadas.

Se multiplicarmos as variâncias $v_i = \mathrm{Var}(y_i)$, $i = 1, \ldots, n$, por qualquer constante, o bias manter-se-á constante. Isto é importante pois reflecte a existência de uma separação entre os erros que se cometem quando se medem a incidência da doença e a exposição.

Assim, deverá multiplicar-se a variância σ_ε^2 por um factor independente da precisão com que se mede a exposição.

Consideremos a função $\sigma_\varepsilon^2 = h(c)$ que traduz a dependência da variância de um dado custo c.

Em geral esta função é uma função decrescente. Se definirmos um máximo aceitável para a variância, $\overline{\sigma}_\varepsilon^2$, conseguiremos determinar um custo mínimo aceitável, $\underline{c}$, para as medições da exposição.

4.3 Limites para o Enviesamento

Para obter estes limites de confiança para $\Delta = \widetilde{\beta}_1 - \widehat{\beta}_1$, vamos utilizar a aproximação de Edgeworth para a distribuição de Δ.

Esta aproximação é dada por

$$F_\Delta(\mathfrak{z}) \approx \Phi(\mathfrak{z}) - \frac{1}{3!\sqrt{n}} \lambda_3 \Phi^{(3)}(\mathfrak{z}) + \frac{1}{n}[\frac{1}{4!}\lambda_4 \Phi^{(4)}(\mathfrak{z}) + \frac{10}{6!}\lambda_3^2 \Phi^{(6)}(\mathfrak{z})]$$

onde $\Phi(\mathfrak{z})$ representa a função densidade da distribuição normal estandardizada, $\Phi^{(j)}(\mathfrak{z})$ a sua derivada de ordem j e λ_j é o cumulante estandardizado, isto é, $\lambda_j = \frac{\kappa_j}{\overset{\circ}{\sigma}^j}$ com $\overset{\circ}{\sigma}$ o desvio padrão e κ_j o j-ésimo cumulante, tendo-se

$$\begin{cases} \kappa_1 = \overset{\circ}{\mu} \\ \kappa_2 = \overset{\circ}{\sigma} = \overset{\circ}{\mu}'_2 - \overset{\circ}{\mu}^2 \\ \kappa_3 = \overset{\circ}{\mu}_3 = \overset{\circ}{\mu}'_3 - 3\overset{\circ}{\mu}\overset{\circ}{\mu}'_2 + 2\overset{\circ}{\mu}^3 \\ \kappa_4 = \overset{\circ}{\mu}_4 - 3(\overset{\circ}{\sigma}^2)^2 = \overset{\circ}{\mu}'_4 - 4\overset{\circ}{\mu}\overset{\circ}{\mu}'_3 - 3\overset{\circ}{\mu}'^2_2 + 10\overset{\circ}{\mu}^2\overset{\circ}{\mu}'_2 - 4\overset{\circ}{\mu}^4 \end{cases}$$

com $\overset{\circ}{\mu}'_j$ o momento de ordem j e $\overset{\circ}{\mu}_j$ o momento central de ordem j.

Sendo

$$\Delta \approx \frac{1}{d}\left[a\left(\tilde{v}-v\right)-\left(\tilde{b}-b\right)u\right]-\frac{\hat{\beta}_1}{d}\left(\tilde{d}-d\right)=\theta-\frac{\hat{\beta}_1}{d}\left(\tilde{d}-d\right)$$

com

$\theta=\frac{1}{d}\left[a\left(\tilde{v}-v\right)-\left(\tilde{b}-b\right)u\right]$começamos por calcular as suas potências

$$\Delta^2 \approx \theta^2-2\theta\frac{\hat{\beta}_1}{d}\left(\tilde{d}-d\right)+\frac{\hat{\beta}_1^2}{d^2}\left(\tilde{d}-d\right)^2$$

$$\Delta^3 \approx \theta^3-3\theta^2\frac{\hat{\beta}_1}{d}\left(\tilde{d}-d\right)+3\theta\frac{\hat{\beta}_1^2}{d^2}\left(\tilde{d}-d\right)^2-\frac{\hat{\beta}_1^3}{d}\left(\tilde{d}-d\right)^3$$

e

$$\Delta^4 \approx \theta^4-4\theta^3\frac{\hat{\beta}_1}{d}\left(\tilde{d}-d\right)+6\theta^2\frac{\hat{\beta}_1^2}{d^2}\left(\tilde{d}-d\right)^2-4\theta\frac{\hat{\beta}_1^3}{d^3}\left(\tilde{d}-d\right)^3+\frac{\hat{\beta}_1^4}{d^4}\left(\tilde{d}-d\right)^4$$

com o objectivo de obter os quatro primeiros momentos de Δ, necessários para a construção dos intervalos. Momentos esses que representamos por

$$\overset{\circ}{\mu}=\mathrm{E}(\Delta),\qquad \overset{\circ}{\mu}{}'_2=\mathrm{E}(\Delta^2),\qquad \overset{\circ}{\mu}{}'_3=\mathrm{E}(\Delta^3),\qquad \overset{\circ}{\mu}{}'_4=\mathrm{E}(\Delta^4).$$

As expressões destes momentos são bastante extensas pelo que as remetemos para apêndice (capítulo 6).

Substituindo-se estes momentos nas expressões de κ_1, κ_2, κ_3 e κ_4 obtém-se a expressão de $F_\Delta(\mathfrak{z})$. Resolvendo numericamente (por exemplo, pelo Método da Bisseção, ver Capítulo 2) o par de equações

$$\begin{cases} F_\Delta(\mathfrak{z})=\frac{\alpha}{2} \\ \\ F_\Delta(\mathfrak{z})=1-\frac{\alpha}{2} \end{cases}$$

obtêm-se os quantis $\mathfrak{z}_{\alpha/2}$ e $\mathfrak{z}_{1-\alpha/2}$, que nos dão os intervalos de confiança aproximados para o valor do enviesamento, $[\mathfrak{z}_{\alpha/2};\mathfrak{z}_{1-\alpha/2}]$.

Capítulo 5

Delineamento de Estudos de Campo

5.1 O Problema

Como referimos no capítulo introdutório, a falta de dados concretos levou-nos a utilizar o método dos Mínimos Quadrados Estruturados para obter medidas aproximadas de exposições. No capítulo anterior vimos que estas conduzem a estimadores enviesados por defeito da taxa de variabilidade com a exposição β_1. Assim, os resultados apresentados nas aplicações têm de ser considerados como limites inferiores.

Surge, pois, a necessidade de vir a dispor de observações mais precisas das incidências. Este problema é de grande interesse para Portugal dadas as taxas elevadas de certas doenças, como a Tuberculose e a Sida. Para evolução desta última veja-se por exemplo Oliveira e Mexia (2004).

Assim, neste capítulo delineamos estudos de campo que esperamos venham a ser realizados. Dado que tais estudos devem ser previamente planeados, a inclusão deste capítulo não nos parece despropositada. Aliás é metodologicamente correcto começar-se por resolver os problemas de planeamento.

No que se segue consideramos, a partir duma partição da variância, vários cenários, sucessivamente mais completos, para a utilização de estações de monitorização de exposições.

5.2 Partição da Variância

Comecemos por relembrar a expressão do valor esperado do bias

$$\mathrm{E}(\widetilde{\beta}_1 - \widehat{\beta}_1) \approx -\beta_1 \frac{[(\sum_{i=1}^{n} \frac{1}{v_i})^2 - \sum_{i=1}^{n} \frac{1}{v_i^2}]}{(\sum_{i=1}^{n} \frac{1}{v_i})(\sum_{i=1}^{n} \frac{f_i^2}{v_i}) - (\sum_{i=1}^{n} \frac{f_i}{v_i})^2} \sigma_\varepsilon^2 < 0 \tag{5.1}$$

onde $f_1, \ldots, f_n$ são os valores exactos das exposições.

Usam-se as medidas aproximadas $\tilde{f}_i = f_i + \varepsilon_i$, $i = 1, \ldots, n$, com o objectivo de obter o estimador $\widetilde{\beta}_1$ para a taxa de variação do logit relativamente à exposição, enquanto $\widehat{\beta}_1$ é o estimador que obteríamos se utilizássemos as medidas exactas.

Convém ainda recordar que a equação (5.1) foi obtida partindo dos pressupostos:

$$\begin{cases} \mathrm{E}(\varepsilon_i) = 0; \\ \mathrm{Var}(\varepsilon_i) = \sigma^2_\varepsilon; \\ \text{independência dos erros.} \end{cases}$$

É importante salientar que o denominador da fracção da equação (5.1) pode ser utilizado como estimador de σ^2_f, que corresponde à componente da variância associada às diferenças existentes nas exposições obtidas nas várias estações de medição.

O próximo passo consistirá em particionar a variância σ^2_ε em duas componentes

$$\sigma^2_\varepsilon = \sigma^2_a + \sigma^2_e$$

onde

- σ^2_a é a componente da variância associada aos erros de amostragem dentro do conjunto populacional associado a cada uma das estações de medição;
- σ^2_e é a componente da variância associada aos erros de medição.

Ao decompormos a variância σ^2_ε na soma de duas componentes estamos a admitir que as duas causas de erro (amostragem e técnica usada) actuam independentemente, o que nos parece um pressuposto perfeitamente aceitável.

Admita-se também que temos n regiões donde se recolhem amostras de dimensão m, obtendo-se os logits

$$y_i = \ln \frac{x_i}{m - x_i} \ , \ i = 1, \ldots, n,$$

com variâncias $v_1, \ldots, v_n$.

Na expressão anterior, os x_i têm distribuição binomial com parâmetro m. Relembremos que p_i, $i = 1, \ldots, n$, representa a probabilidade de um indivíduo pertencente à i-ésima região estar infectado.

Admitindo que $1 - p_i \approx 1$, $i = 1, \ldots, n$, obtemos

$$\tfrac{dy}{dx} = \tfrac{m}{x_i(m-x_i)} \approx \tfrac{1}{p_i(1-p_i)} \approx \tfrac{1}{p_i} \ , \ i = 1, \ldots, n,$$

e consequentemente

$$v_i \approx \tfrac{1}{p_i^2} \tfrac{p_i}{m} = \tfrac{1}{mp_i} \ , \ i = 1, \ldots, n.$$

Por outro lado, sabemos que

$$p_i = \frac{e^{\beta_0+\beta_1 f_i}}{1+e^{\beta_0+\beta_1 f_i}} \approx e^{\beta_0+\beta_1 f_i} \ , \ i = 1, \ldots, n,$$

e assim

$$\frac{1}{1+e^{\beta_0+\beta_1 f_i}} = 1 - p_i \approx 1 \ , \ i = 1, \ldots, n,$$

pelo que

$$v_i \approx \frac{e^{-\beta_0-\beta_1 f_i}}{m}.$$

Finalmente, obtemos

$$\mathbb{K} = \frac{(\sum_{i=1}^{n} \frac{1}{v_i})^2 - \sum_{i=1}^{n} \frac{1}{v_i^2}}{(\sum_{i=1}^{n} \frac{1}{v_i})(\sum_{i=1}^{n} \frac{f_i^2}{v_i}) - (\sum_{i=1}^{n} \frac{f_i}{v_i})^2} \approx \frac{(\sum_{i=1}^{n} e^{\beta_1 f_i})^2 - \sum_{i=1}^{n} e^{2\beta_1 f_i}}{(\sum_{i=1}^{n} e^{\beta_1 f_i})(\sum_{i=1}^{n} e^{\beta_1 f_i} f_i^2) - (\sum_{i=1}^{n} e^{\beta_1 f_i} f_i)^2}$$

o que mostra que apenas β_1 e $\boldsymbol{f}^n$, o vector cujas componentes são os valores exactos das exposições, influenciam significativamente $\mathbb{K}$.

Por outro lado, uma vez que p_i, $i = 1, \ldots, n$, são significativamente inferiores a 1 ($p_i \ll 1$) e $v_i \approx \frac{1}{mp_i}$, $i = 1, \ldots, n$, temos que

$$\sum_{i=1}^{n} \frac{1}{v_i^2} \ll (\sum_{i=1}^{n} \frac{1}{v_i})^2 \ .$$

De acordo com estes resultados, vê-se que

$$\mathbb{K} \approx \frac{\sum_{i=1}^{n} \frac{1}{v_i}}{\sum_{i=1}^{n} \frac{f_i^2}{v_i} - \frac{(\sum_{i=1}^{n} \frac{f_i}{v_i})^2}{\sum_{i=1}^{n} \frac{1}{v_i}}} = \frac{\sum_{i=1}^{n} p_i}{\sum_{i=1}^{n} p_i f_i^2 - \frac{(\sum_{i=1}^{n} p_i f_i)^2}{\sum_{i=1}^{n} p_i}} = \frac{1}{\frac{\sum_{i=1}^{n} p_i (f_i - f_0)^2}{\sum_{i=1}^{n} p_i}}$$

com

$$\begin{cases} f_0 = \dfrac{\sum_{i=1}^{n} p_i f_i}{\sum_{i=1}^{n} p_i} = \sum_{i=1}^{n} q_i f_i \\ \\ q_i = \dfrac{p_i}{\sum_{j=1}^{n} p_j}, i = 1, \ldots, n. \end{cases}$$

Obtemos, desta forma, o resultado

$$\mathrm{E}(\widetilde{\beta}_1 - \widehat{\beta}_1) \approx -\beta_1 \frac{\sigma_a^2 + \sigma_e^2}{\sigma_f^2}$$

onde

$$\sigma_f^2 = \sum_{i=1}^{n} q_i (f_i - f_0)^2.$$

5.3 Cenários

Apresentamos agora vários cenários possíveis.

Primeiro Cenário

As estações e as sub-populações a considerar estão previamente escolhidas. Como cada estação estará adstrita a uma região, tanto as estações como as regiões serão conhecidas. Das três componentes da variância a única sobre a qual se pode actuar é σ_e^2. Admitamos que existe uma relação custos/precisão para esta componente, dada por uma função decrescente e com uma assímptota horizontal, como por exemplo a que apresentamos na Figura (5.1).

O objectivo consiste em determinar o ponto C' de custo por estação, a partir do qual os ganhos de precisão deixam de ser relevantes, quer isto dizer que à direita deste ponto há apenas uma diminuição limitada na variância σ_e^2. Convém salientar que, neste cenário com $\overline{\sigma}_e^2$ o valor de σ_e^2 para $C = C'$, será difícil ter

$$|\mathrm{E}(\widetilde{\beta}_1 - \widehat{\beta}_1)| < \beta_1 \frac{\sigma_a^2 + \overline{\sigma}_e^2}{\sigma_f^2}.$$

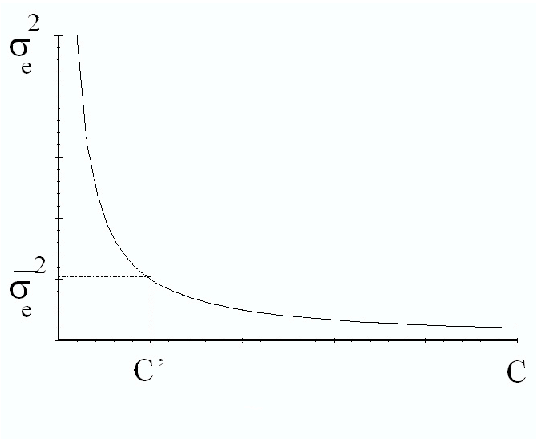

Figura 5.1: Relação Custos vs Precisão.

Com o objectivo de reduzir o mais possível σ_ε^2, podemos sub-dividir a população em sub-populações, isto é, proceder a uma estratificação. No entanto os custos associados às medições deverão, em princípio, ser idênticos para as diferentes situações e, uma vez que p_i, $i = 1, \ldots, n$, é muito inferior a 1, a mesma estabilidade existe quanto às variâncias v_i, $i = 1, \ldots, n$. Assim, é aceitável que todas as amostras tenham dimensão m.

O custo total poderá então ser dado pela expressão

$$C = c(n) + c_1 n + c_2 nm \tag{5.2}$$

onde $c(n)$ é um custo fixo dependente apenas do número de estações, c_1 um custo por estação e c_2 um custo por pessoa examinada.

Podemos admitir que a variância σ_a^2 é inversamente proporcional à dimensão das amostras, isto é, $\sigma_a^2 = \frac{a}{m}$. Admitindo a já referida relação entre custos/precisão através duma função decrescente, podemos escrever $\sigma_e^2 = \eth(c_1)$, com $\eth'(c_1) < 0$ e $\eth''(c_1) > 0$.

Pressupondo ainda que conhecemos n e sendo $A = C - c(n)$, pretendemos minimizar

$$\sigma_a^2 + \sigma_e^2 = \frac{a}{m} + \eth(c_1).$$

Reescreva-se a equação (5.2) na forma

$$c_1 + c_2 m = \frac{A}{n}$$

uma vez que, sendo n dado, $B = \frac{A}{n}$ é conhecido.

Usando a técnica dos multiplicadores de Lagrange[1] construímos a função auxiliar

$$\eth(c_1, m, \lambda) = \tfrac{a}{m}\eth(c_1) - \lambda(c_1 + c_2 m - B)$$

onde se considera que c_2 também é conhecido.

Somos, assim, levados a resolver o sistema

$$\begin{cases} \frac{\partial \eth}{\partial c_1} = \eth'(c_1) - \lambda = 0 \\ \\ \frac{\partial \eth}{\partial m} = -\frac{a}{m^2} - \lambda c_2 = 0 \\ \\ c_1 + c_2 m - B = 0. \end{cases}$$

Uma vez especificada a função $\partial\eth$, que dependerá do problema em estudo, teremos que recorrer aos métodos numéricos para obter a solução deste sistema.

Segundo Cenário

As estações estão implantadas mas os limites das regiões a atribuir-lhes não estão definidos.

---

[1] O método dos multiplicadores de Lagrange é utilizado para encontrar os extremos de uma dada função multivariada

$$f(x_1, x_2, \ldots, x_n)$$

sob a condição

$$g(x_1, x_2, \ldots, x_n) = 0,$$

onde f e g são funções contínuas com derivadas parciais de primeira ordem pertencentes ao conjunto aberto que contém a curva $g(x_1, x_2, \ldots, x_n) = 0$, e $\nabla_g \neq 0$ para qualquer ponto dessa curva (onde ∇ representa o gradiente). Para que um extremo exista,

$$df = \frac{df}{dx_1}dx_1 + \frac{df}{dx_2}dx_2 + \cdots + \frac{df}{dx_n}dx_n = 0.$$

Como g é constante, temos que

$$dg = \frac{dg}{dx_1}dx_1 + \frac{dg}{dx_2}dx_2 + \cdots + \frac{dg}{dx_n}dx_n = 0.$$

Então, se multiplicarmos esta segunda equação por um parâmetro λ e lhe adicionarmos a primeira, obtemos

$$(\frac{df}{dx_1} + \lambda\frac{dg}{dx_1})dx_1 + (\frac{df}{dx_2} + \lambda\frac{dg}{dx_2})dx_2 + \cdots + (\frac{df}{dx_n} + \lambda\frac{dg}{dx_n})dx_n = 0.$$

Portanto,

$$\frac{df}{dx_k} + \lambda\frac{dg}{dx_k} = 0$$

para todo o $k = 1, 2, \ldots, n$. A constante λ designa-se por multiplicador de Lagrange.

No cenário anterior, considerámos que $\sigma_a^2 = \frac{a}{m}$, não nos podemos esquecer que se trata de uma expressão aproximada. Suponhamos agora que a população atribuída a uma estação é constituída por L sub-populações, tendo-se desta forma

$$a = a_1 + a_2$$

onde a_1 mede a variação interna das sub-populações e a_2 mede a variação entre sub-populações.

Uma vez que temos n populações, iremos ter outras tantas decomposições. Por outro lado, a cada estação estará associada uma sub-população para a qual ela é mais representativa, a sub-população padrão.

Neste cenário é possível influenciar (minimizar) ambas as componentes da variância σ_a^2 e σ_e^2, já que se pode variar a atribuição das sub-populações. Deste modo conseguimos o controlo de σ_ε^2 e, consequentemente, do bias.

Uma regra que se pode tentar usar é atribuir cada sub-população à estação de cuja sub-população padrão ela se aproxima mais. Para realizar esta tarefa há que:

- escolher as sub-populações padrão associadas às diferentes estações;
- definir uma medida de dissemelhança[2] entre populações;
- construir a matriz de dissemelhanças.

[2]Uma medida de dissemelhança é uma função, d, que a cada par de indivíduos faz corresponder um valor de um espaço euclidiano unidimensional (normalmente R). De um modo geral, estas medidas tomam valores no intervalo [0,1]. Os valores das medidas de dissemelhança não são, na maioria dos casos, observados directamente, mas sim calculados a partir de uma matriz de dados, designada por matriz de dissemelhanças.

Sendo d_{ij} a medida de dissemelhança entre os indivíduos i, $i = 1, \ldots, n$, e j, $j = 1, \ldots, p$, a função d deve verificar as seguintes propriedades:

1. $d_{ij} \geq 0, \forall_{i,j}$;
2. $d_{ii} = 0, \forall_i$;
3. $d_{ij} = d_{ji}, \forall_{i,j}$;
4. $d_{ir} + d_{rj} \geq d_{ij}, \forall_{i,j,r}$;
5. $d_{ij} = 0$ se e só se $i = j, \forall_{i,j}$.

Se d verificar as quatro primeiras propriedades, chama-se semi-métrica; se verificar as cinco propriedades, denomina-se métrica (distância) e, se verificar as três primeiras e ainda a propriedade $d_{ij} \leq \max(d_{ir}, d_{rj}), \forall_{i,j,r}$, denomina-se ultramétrica.

Podem definir-se várias medidas de dissemelhança entre objectos. Vamos, de uma forma sucinta, enumerar as mais comuns:

- Distância Euclidiana - $d_{ij} = [(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)]^{1/2}$, onde $\boldsymbol{x}_i$ representa o vector das observações do indivíduo i. Esta distância, embora seja a mais utilizada, apresenta alguns problemas: não é invariante às mudanças de escala, tem um comportamento irregular quando as variáveis têm variâncias muito diferentes e o mesmo acontece quando as variáveis estão correlacionadas.
- Quadrado da Distância Euclidiana, a única vantagem em relação à anterior reside no facto de ter um melhor desempenho quando as variáveis são correlacionadas.

Terceiro Cenário

O terceiro cenário é um cenário livre.

Suponhamos que apenas se tem uma ideia aproximada do número de estações a implantar. Podemos começar a partir da matriz de dissemelhanças considerada no cenário anterior e aplicar uma análise de clusters procurando identificar as sub-populações que funcionam como centros de gravidade. Serão então estas as sub-populações escolhidas como sub-populações padrão associadas às estações de medição. Poderá acontecer que nem todas as estações candidatas possam ser implantadas.

É importante salientar que, neste cenário é possível actuar ao nível das três componentes da variância: σ_a^2, σ_e^2 e σ_f^2. Desta forma conseguiremos um melhor controlo do bias do que nos dois primeiros cenários.

O nosso objectivo é começar por minimizar

$$\mathbb{K}(\sigma_a^2 + \sigma_e^2) = \frac{\sigma_a^2+\sigma_e^2}{\sigma_f^2} \ .$$

Admitindo que se estão a comparar cenários para os quais os valores de $\sigma_a^2 + \sigma_e^2$ são similares, poderemos então, numa segunda fase, minimizar

$$\sigma_f^2 = \sum_{i=1}^{n} q_i (f_i - f_0)^2 \ .$$

Observemos ainda que, se tivermos L possíveis sub-populações das quais se escolhem l sub-populações padrão, o número de escolhas é $\begin{pmatrix} L \\ l \end{pmatrix}$.

Informação Necessária

Para se poder actuar ao nível do segundo e terceiro cenários é necessário dispor de informação que permita obter a matriz de dissemelhanças.

Para obter tal informação deveremos começar por elaborar uma lista de variáveis que determinem a exposição e caracterizem as populações. O ideal seria:

- Distância Euclidiana Estandardizada (Distância de Karl Pearson) - $d_{ij} = [(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{D}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)]^{1/2}$, onde $\boldsymbol{D}$ representa a matriz diagonal das variâncias das variáveis. Continua a ser invariante às mudanças de escala.
- Distância Euclidiana Média - $d_{ij} = [(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \frac{1}{p} \boldsymbol{I}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)]^{1/2}$, a única vantagem relativamente à distância Euclidiana está no seu melhor desempenho quando existem dados omissos.
- Distância de Mahalanobis - $d_{ij} = [(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)]^{1/2}$, onde $\boldsymbol{S}$ representa a matriz de covariâncias das variáveis. Esta distância resolve ambos os problemas, da invariância de escalas distintas e da correlação entre as variáveis.

1. fazer o levantamento das variáveis na população;
2. definir as sub-populações;
3. construir a matriz de dissemelhanças;
4. escolher as sub-populações candidatas a sub-populações padrão;
5. seleccionar as populações padrão e os limites que definem cada uma das regiões;
6. proceder à implantação das estações de medição;
7. proceder ao processo de amostragem em cada uma das regiões.

No final deste processo, a partir dos valores ajustados, deveremos conseguir reconstituir a incidência média $\widetilde{p}$.

Por outro lado, as variáveis escolhidas para definir as populações devem incluir a residência e/ou local de trabalho para facilitar a delimitação das regiões.

Convém observar que existe uma certa semelhança entre o método que estamos a tentar desenvolver e o "método de tarifação à priori". Assim, podemos partir duma lista preliminar e extensa de variáveis discretas, verificando quais destas estão significativamente associadas a variações nas incidências. Segue-se a definição das sub-populações a partir das variáveis seleccionadas. Deve-se repetir este processo até se ter obtido o conjunto "óptimo"de variáveis.

Quarto Cenário

Consideremos ainda um último cenário.

Suponhamos que, neste quarto cenário, o número de estações, n, não é conhecido.

Já vimos que, com

$$\Delta = \widetilde{\beta}_1 - \widehat{\beta}_1$$

tem-se

$$\mathrm{E}(\Delta) \approx -\beta_1 \sigma_\varepsilon^2 \frac{\Delta(V^{-1})}{\Delta(f)}$$

onde

$$\begin{cases} \sigma_\varepsilon^2 = \sigma_a^2 + \sigma_e^2 \leq \frac{\mathbb{K}}{m} + \sigma_e^2 \\ \Delta(V^{-1}) = -[\sum_{i=1}^{n} \frac{1}{v_i^2} - (\sum_{i=1}^{n} \frac{1}{v_i})^2] \\ \Delta(f) = (\sum_{i=1}^{n} \frac{1}{v_i})(\sum_{i=1}^{n} \frac{f_i^2}{v_i}) - (\sum_{i=1}^{n} \frac{f_i}{v_i})^2 \end{cases}$$

com $v_i \approx \frac{m}{p_i}$, $p_i \approx e^{\beta_0+\beta_1 f_i}$ e m a dimensão das amostras.

Recordemos ainda que, para o valor do custo total, temos a expressão

$$C = c + c_1 n + c_2 nm.$$

É, então, fácil verificar que

$$\frac{\Delta(V^{-1})}{\Delta(f)} = \breve{\eth}(\boldsymbol{f}^n) \ .$$

Vimos também que σ_e^2 será dada por uma função do custo por indivíduo examinado, c_2, sendo possível assumir que se tem uma curva idêntica à apresentada na Figura (5.2).

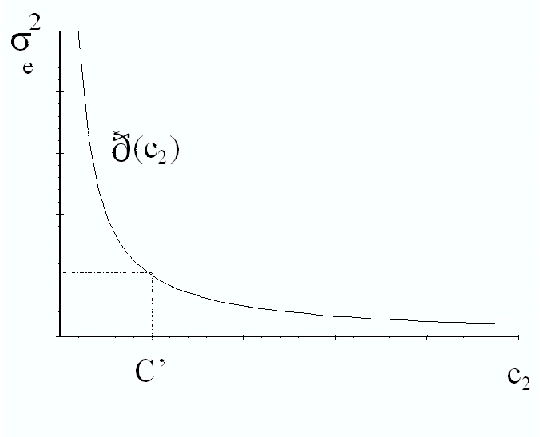

Figura 5.2: Relação Custos vs Precisão.

Esta curva tem uma assímptota horizontal, o que sugere a existência dum valor C' para além do qual não existe interesse em continuar.

Observemos ainda que

$$ln|\mathrm{E}(\Delta)| \leq ln\beta_1| + ln(\tfrac{\mathbb{K}}{m} + \eth(c_2)) + ln(\breve{\eth}(\boldsymbol{f}^n))$$

sendo que no segundo membro desta desigualdade se tem:

- um factor constante, $ln|\beta_1|$;
- um factor dependente de m e c_2;

- um factor dependente de $\boldsymbol{f}^n$.

Uma vez que o nosso objectivo consiste em minimizar $|\mathrm{E}(\Delta)|$, podemos começar por tentar obter o mínimo de

$$\frac{\Delta(V^{-1})}{\Delta(f)} = \breve{\eth}(\boldsymbol{f}^n)$$

em função de n.

Para atingirmos este objectivo, podemos proceder da seguinte forma:

1. começar com uma lista de variáveis que influenciam a exposição;
2. proceder a um estudo cuidado das variáveis seleccionadas para testar a sua significância;
3. usar as variáveis seleccionadas para definir N sub-populações;
4. usar os resultados disponíveis para estimar a exposição $\ddot{f}_1, \ldots, \ddot{f}_N$, associada a cada uma das N sub-populações;
5. para $l = 2, \ldots, N$ escolher l sub-populações com o objectivo de minimizar $\breve{\eth}(\boldsymbol{f}^n)$.

Como $\breve{\eth}(\boldsymbol{f}^n)^{-1}$ é similar a uma soma de quadrados dos resíduos para a média, no passo 5. devemos escolher populações extremas.

Sendo m_n o mínimo de $ln(\breve{\eth}(\boldsymbol{f}^n))$, o problema pode ser reformulado através da minimização da função

$$\mathbb{L} = ln|\beta_1| + ln(\tfrac{\mathbb{K}}{m} + \eth(c_2)) + m_n$$

sob a restrição dada pelo custo total que estamos preparados para assumir.

5.4 Esquema de Implementação - Uma Aplicação

Admita-se que existem diferentes "tipos"de cidades onde se pretende instalar as estações de medição ambiental.

Existem cidades que será necessário subdividir em vários níveis e cidades (mais pequenas) para as quais não haverá necessidade de fazer qualquer subdivisão, pois a própria cidade constitui um único nível.

Esquematicamente ter-se-á algo com uma forma semelhante à apresentada na Figura 5.3.

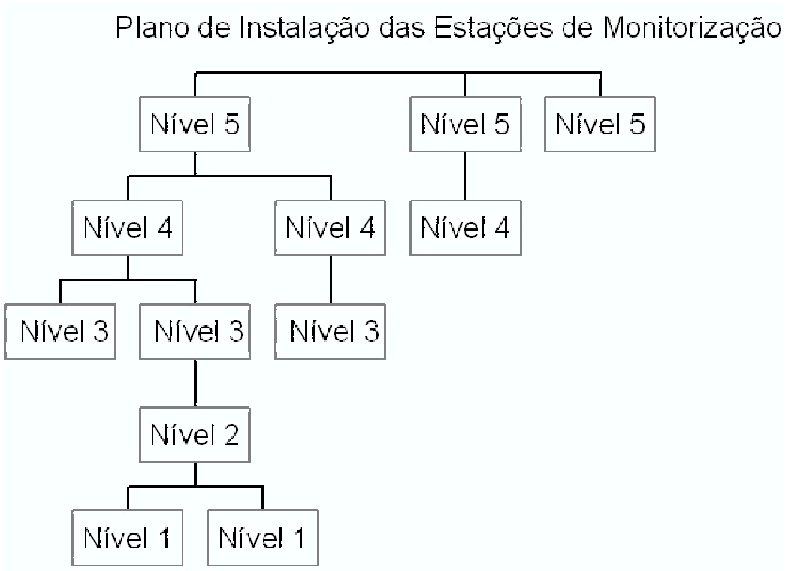

Figura 5.3: Diagrama de Instalação das Estações de Monitorização.

A ideia base consiste em testar a identidade das exposições começando no nível mais baixo e avançando até atingir o nível mais elevado (no hipotético exemplo apresentado o nível 1 corresponde ao nível mais baixo e o nível 5 ao mais elevado).

Pretende-se proceder à instalação das estações de monitorização (medição ambiental) de acordo com a identidade existente entre as exposições, tentando deste modo:

- evitar a instalação de estações em locais desnecessários;
- garantir que locais (regiões) onde não exista concordância entre exposições sejam, garantidamente, contemplados com uma estação de monitorização.

O objectivo principal é minimizar o custo na instalação das estações de monitorização.

A forma de determinar os locais correctos para a instalação das várias estações tem por base os testes de razão de verosimilhanças. Mais concretamente, deverá haver uma estação de monitorização por cada região maximal de verosimilhança com exposições idênticas.

Capítulo 6

Ideias Futuras

A falta de dados limita, sem dúvida, qualquer trabalho, pelo que um trabalho a realizar num futuro muito próximo será aplicar a metodologia desenvolvida a um conjunto de dados em que medições de impactos ambientais e a sua influência na saúde estejam contempladas, podendo deste modo aperfeiçoar o trabalho desenvolvido.

O que mais nos interessa será poder aplicar os nossos resultados em Estudos de Campo, pondo em prática todas as "guidelines"que sugerimos no Capítulo 5 desta dissertação.

Sabemos que, por si só, estas "guidelines"serão insuficientes para elaborar um estudo completo na área da Epidemiologia Ambiental, no entanto podemos adiantar, de forma breve, ideias que julgamos indispensáveis num estudo de campo desta natureza.

A informação sobre a exposição e efeitos subsequentes traduzem a realidade no delineamento dos estudos epidemiológicos. É esta informação que influencia decisivamente o tipo de estudo a ser desenvolvido, qual a sua duração, a escolha da população alvo, as variáveis a serem estudadas, a elaboração de instrumentos de pesquisa e o método de recolha das amostras.

Para avaliar a exposição, as variáveis definidas devem responder a questões básicas para as investigações epidemiológicas, de que são exemplo:

- quais os indivíduos ou grupos de indivíduos mais expostos;
- quais os locais onde estão situadas as fontes de poluição e que características destes locais podem interferir na exposição;
- qual a frequência, magnitude e duração da exposição;
- quais as diferentes vias de absorção e respectivos riscos.

Esta tarefa, como já referimos, é dificultada pela ausência ou fragilidade dos registos destas informações e pelo facto de abranger diversos sectores, não sendo fácil estabelecer a ligação entre os mesmos.

No entanto, existem outras dificuldades inerentes ao facto de estarmos a relacionar a saúde e o ambiente, por exemplo, a população de interesse abrange toda a população e todos os grupos populacionais, isto é, indivíduos de qualquer idade, sexo,

ocupação, condição sócio-económica, estado de saúde, etc. Por se tratar de ambiente, também as características sociais e físicas do local são importantes, além de outros factores como condições meteorológicas, topográficas, hidrográficas e geológicas. A identificação das fontes de emissão dos poluentes é dificultada pela sua variabilidade e dispersão. Para não falar de todos os problemas cuja resposta é obtida recorrendo à Toxicologia. É muitas vezes difícil diagnosticar intoxicações por poluentes químicos, pois não existe um quadro clínico "clássico"para a maioria das substâncias químicas. Além disso, a exposição não está relacionada com um único poluente mas sim com uma multiplicidade deles, que pode ou não provocar sinais e sintomas diferentes.

Relacionado com as aplicações apresentadas no Capitulo 3 surgiu o seguinte problema:

- os valores da taxa de variação, β_1, tendem a ser maiores quando se consideram todos os tipos de Tuberculose do que quando apenas se considera apenas a Tuberculose Pulmonar.

É verdade que, com os dados de que dispomos, uma simples comparação não é completamente fiável. No entanto não conseguimos justificar estes resultados, pelo que pretendemos vir a responder a esta questão com a ajuda de novos dados.

Como podemos constatar, muito é o trabalho que nos espera no futuro, pois encaramos uma tese de doutoramento como a obtenção da "carta de condução"para que, a partir da mesma, se possa desenvolver um trabalho cada vez mais completo e pormenorizado.

Apêndice A

Momentos de $\widehat{\beta}_1$

Comecemos por apresentar as expressões de $\widehat{\beta}_1$ e das suas potências.

$$\widehat{\beta}_1 = \frac{1}{d}\left[a\sum_{i=1}^{n}\frac{f_i\ y_i}{v_i} - \sum_{i=1}^{n}\frac{y_i}{v_i}\right] = \frac{1}{d}\sum_{i=1}^{n}\frac{z_i\ y_i}{v_i}$$

com $z_i = af_i - b$

$$\widehat{\beta}_1^2 = \frac{1}{d^2}\left[\sum_{i=1}^{n}\frac{z_i^2\ y_i^2}{v_i^2} + \sum_{\substack{i=1\\ i\neq l}}^{n}\sum_{l=1}^{n}\frac{z_i\ z_l\ y_i\ y_l}{v_i\ v_l}\right]$$

$$\widehat{\beta}_1^3 = \frac{1}{d^3}\left[\left(\sum_{i=1}^{n}\frac{z_i\ y_i}{v_i}\right)\left(\sum_{j=1}^{n}\frac{z_j^2\ y_j^2}{v_j^2}\right) + \right.$$

$$\left. + \left(\sum_{i=1}^{n}\frac{z_i\ y_i}{v_i}\right)\left(\sum_{\substack{j=1\\ j\neq l}}^{n}\sum_{l=1}^{n}\frac{z_j\ z_l\ y_j\ y_l}{v_j\ v_l}\right)\right] =$$

$$= \frac{1}{d^3}\left[\sum_{i=1}^{n}\frac{z_i^3\ y_i^3}{v_i^3} + 3\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i^2\ z_j\ y_i^2\ y_j}{v_i^2\ v_j} + \sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_i\ z_j\ z_l\ y_i\ y_j\ y_l}{v_i\ v_j\ v_l}\right]$$

$$\widehat{\beta}_1^4 = \left[\frac{1}{d^4}\left(\sum_{i=1}^{n}\frac{z_i\ y_i}{v_i}\right)\left(\sum_{j=1}^{n}\frac{z_j^3\ y_j^3}{v_j^3}\right) + 3\left(\sum_{i=1}^{n}\frac{z_i\ y_i}{v_i}\right)\left(\sum_{\substack{j=1\\ j\neq l}}^{n}\sum_{l=1}^{n}\frac{z_j^2\ z_l\ y_j^2\ y_l}{v_j^2\ v_l}\right) + \right.$$

$$+\left(\sum_{i=1}^{n}\frac{z_i\ y_i}{v_i}\right)\left(\sum_{\substack{j=1\\ j\neq l\neq k}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{z_j\ z_l\ z_k\ y_j\ y_l\ y_k}{v_j\ v_l\ v_k}\right)\Bigg] =$$

$$= \frac{1}{d^4}\left[\sum_{i=1}^{n}\frac{z_i^4\ y_i^4}{v_i^4} + 4\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i^3\ z_j\ y_i^3\ y_j}{v_i^3\ v_j} + 3\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i^2\ z_j^2\ y_i^2\ y_j^2}{v_i^2\ v_j^2} + \right.$$

$$\left. +6\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_i^2\ z_j\ z_l\ y_i^2\ y_j\ y_l}{v_i^2\ v_j\ v_l} + \sum_{\substack{i=1\\ i\neq j\neq l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{z_i\ z_j\ z_l\ z_k\ y_i\ y_j\ y_l\ y_k}{v_i\ v_j\ v_l\ v_k}\right]$$

$$\mathrm{E}\left[\widehat{\beta}_1\right] = \tfrac{1}{d}\sum_{i=1}^{n}\frac{z_i\ \mu_i}{v_i}$$

$$\mathrm{E}\left[\widehat{\beta}_1^2\right] = \tfrac{1}{d^2}\left(\sum_{i=1}^{n}\frac{z_i^2\ \mu'_{i,2}}{v_i^2} + \sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i\ z_j\ \mu_i\ \mu_j}{v_i\ v_j}\right)$$

$$\mathrm{E}\left[\widehat{\beta}_1^3\right] = \tfrac{1}{d^3}\left(\sum_{i=1}^{n}\frac{z_i^3\ \mu'_{i,3}}{v_i^3} + 3\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i^2\ z_j\ \mu'_{i,2}\ \mu_j}{v_i^2\ v_j} + \sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_i\ z_j\ z_l\ \mu_i\ \mu_j\ \mu_l}{v_i\ v_j\ v_l}\right)$$

$$\mathrm{E}\left[\widehat{\beta}_1^4\right] = \tfrac{1}{d^4}\left(\sum_{i=1}^{n}\frac{z_i^4\ \mu'_{i,4}}{v_i^4} + 4\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i^3\ z_j\ \mu'_{i,3}\ \mu_j}{v_i^3\ v_j} + 3\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i^2\ z_j^2\ \mu'_{i,2}\ \mu'_{j,2}}{v_i^3\ v_j} + \right.$$

$$\left. +6\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_i^2\ z_j\ z_l\ \mu'_{i,2}\ \mu_j\ \mu_l}{v_i\ v_j\ v_l} + \sum_{\substack{i=1\\ i\neq j\neq l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{z_i\ z_j\ z_l\ z_k\ \mu_i\ \mu_j\ \mu_l\ \mu_k}{v_i\ v_j\ v_l\ v_k}\right)$$

Apêndice B

Momentos de Δ

Com cálculos algo extensos (ver Volume 2), obtemos os momentos de Δ que intervêm na construção dos intervalos de confiança para o enviesamento.

É importante referir que neste apêndice se apresenta apenas o resultado final desses cálculos, para cada um dos momentos. Para que que isso fosse possível foi necessário recorrer a resultados como a fórmula de Faa di Bruno (ver Capítulo 2).

B.1 Primeiro Momento

$$\begin{aligned}\mathrm{E}\left(\Delta\right) = -\frac{\sigma^2}{d^2}\Bigg[&a^2\sum_{i=1}^{n}\frac{f_i}{v_i^2}\mu_i - a\sum_{i=1}^{n}\frac{f_i}{v_i^3}\mu_i - ab\sum_{i=1}^{n}\frac{1}{v_i^2}\mu_i + b\sum_{i=1}^{n}\frac{1}{v_i^3}\mu_i + \\ &+\mathrm{a}^2\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i}{v_i\ v_j}\mu_i - a\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i}{v_i\ v_j^2}\mu_i - ab\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{1}{v_i\ v_j}\mu_i + \\ &+b\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{1}{v_i\ v_j^2}\mu_i\Bigg]\end{aligned} \tag{B.1}$$

B.2 Segundo Momento

$$\mathrm{E}\left(\Delta^2\right) \approx E\left(\theta^2\right) + E\left(-2\theta\frac{\hat{\beta}_1}{d}\left(\tilde{d}-d\right)\right) + E\left(\frac{\hat{\beta}_1^2}{d^2}\left(\tilde{d}-d\right)^2\right) \tag{B.2}$$

onde

$$\mathrm{E}\left[\theta^2\right] = \frac{\sigma^2}{d^2}\Bigg[a^2\sum_{i=1}^{n}\frac{1}{v_i^2}\mu'_{i,2} - 2a\sum_{i=1}^{n}\frac{1}{v_i^3}\mu'_{i,2} - 2a\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{1}{v_i^2\ v_j}\mu_i\ \mu_j + \sum_{i=1}^{n}\frac{1}{v_i^4}\mu'_{i,2} +$$

$$+2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{1}{v_i^3\ v_j}\mu_i\mu_j+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{1}{v_i^2\ v_j^2}\mu'_{i,2}+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{1}{v_i^2\ v_j\ v_l}\mu_j\mu_l\Bigg]$$

$$\mathrm{E}\left[-2\theta\frac{\hat{\beta}_1}{d}\left(\tilde{d}-d\right)\right]=\frac{-2\sigma^2}{d^3}\left[2a^3\sum_{i=1}^{n}\frac{f_i\ z_i}{v_i^3}\mu'_{i,2}-2a\sum_{i=1}^{n}\frac{f_i\ z_i}{v_i^4}\mu'_{i,2}+\right.$$

$$+2a^3\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i\ z_j}{v_i^2\ v_j}\mu_i\mu_j-2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i\ z_j}{v_i^3\ v_j}\mu_i\mu_j-2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i\ z_i}{v_i^3\ v_j}\mu_i\mu_j-$$

$$\left.-2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i\ z_j}{v_i^2\ v_j^2}\mu'_{j,2}-2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{f_i\ z_l}{v_i^2\ v_j\ v_l}\mu_j\mu_l\right]$$

e

$$\mathrm{E}\left(\frac{\hat{\beta}_1^2}{d^2}\left(\tilde{d}-d\right)^2\right)=\frac{1}{d^4}\left\{\sigma^2\left[4a^2\sum_{i=1}^{n}\frac{f_i^2\ z_i^2}{v_i^4}\mu'_{i,2}+4a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i^2\ z_j^2}{v_i^2\ v_j^2}\mu'_{j,2}+\right.\right.$$

$$\left.+4a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i^2\ z_i\ z_j}{v_i^3\ v_j}\mu_i\mu_j+4a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_i\ f_j^2\ z_j}{v_i\ v_j^3}\mu_i\mu_j+4a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{f_i^2\ z_j\ z_l}{v_i^2\ v_j\ v_l}\mu_j\mu_l\right]+$$

$$+\sigma^4\left[3a^2\sum_{i=1}^{n}\frac{z_i^2}{v_i^4}\mu'_{i,2}-6a\sum_{i=1}^{n}\frac{z_i^2}{v_i^5}\mu'_{i,2}+3\sum_{i=1}^{n}\frac{z_i^2}{v_i^6}\mu'_{i,2}+3a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_j^2}{v_i^2\ v_j^2}\mu'_{j,2}+\right.$$

$$+a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_i^2}{v_i^3\ v_j}\mu'_{i,2}+a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_j^2}{v_i\ v_j^3}\mu'_{j,2}-6a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_j^2}{v_i^3\ v_j^2}\mu'_{j,2}-2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_i^2}{v_i^4\ v_j}\mu'_{i,2}-$$

$$-2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_j^2}{v_i^2\ v_j^3}\mu'_{j,2}+3\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_j^2}{v_i^4\ v_j^2}\mu'_{j,2}+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_i^2}{v_i^4\ v_j^2}\mu'_{i,2}+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_j^2}{v_i^2\ v_j^4}\mu'_{j,2}+$$

$$+2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_i^2}{v_i^4\ v_j}\mu'_{i,2}+2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_j^2}{v_i^2\ v_j^3}\mu'_{j,2}+3a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_i\ z_j}{v_i^3\ v_j}\mu_i\mu_j+3a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_i\ z_j}{v_i\ v_j^3}\mu_i\mu_j+$$

$$+2a^2\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i\ z_j}{v_i^2\ v_j^2}\mu_i\mu_j - 6a\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i\ z_j}{v_i^4\ v_j}\mu_i\mu_j - 6a\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i\ z_j}{v_i\ v_j^4}\mu_i\mu_j -$$

$$-4a\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i\ z_j}{v_i^3\ v_j^2}\mu_i\mu_j + 3\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i\ z_j}{v_i^5\ v_j}\mu_i\mu_j + 3\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i\ z_j}{v_i\ v_j^5}\mu_i\mu_j + 2\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i\ z_j}{v_i^3\ v_j^3}\mu_i\mu_j +$$

$$+4\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{z_i\ z_j}{v_i^3\ v_j^2}\mu_i\mu_j + a^2\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_l^2}{v_i\ v_j\ v_l^2}\mu'_{l,2} - 2a\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_l^2}{v_i^2\ v_j\ v_l^2}\mu'_{l,2} +$$

$$+\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_l^2}{v_i^2\ v_j^2\ v_l^2}\mu'_{l,2} + 2\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_l^2}{v_i^2\ v_j\ v_l^2}\mu'_{l,2} + 3a^2\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_j\ z_l}{v_i^2\ v_j\ v_l}\mu_j\mu_l +$$

$$+4a^2\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_i\ z_l}{v_i^2\ v_j\ v_l}\mu_i\mu_l - 6a\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_j\ z_l}{v_i^3\ v_j\ v_l}\mu_j\mu_l - 4a\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_j\ z_l}{v_i^2\ v_j^2\ v_l}\mu_j\mu_l -$$

$$-4a\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_i\ z_l}{v_i^3\ v_j\ v_l}\mu_i\mu_l + 3\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_j\ z_l}{v_i^4\ v_j\ v_l}\mu_j\mu_l + 4\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_i\ z_l}{v_i^3\ v_j^2\ v_l}\mu_i\mu_l +$$

$$+4\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_j\ z_l}{v_i^2\ v_j^2\ v_l}\mu_j\mu_l + 4\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_i\ z_l}{v_i^3\ v_j\ v_l}\mu_i\mu_l + a^2\sum_{\substack{i=1\\i\neq j\neq l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{z_l\ z_k}{v_i\ v_j\ v_l\ v_k}\mu_l\mu_k -$$

$$-2a\sum_{\substack{i=1\\i\neq j\neq l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{z_l\ z_k}{v_i^2\ v_j\ v_l\ v_k}\mu_l\mu_k + \sum_{\substack{i=1\\i\neq j\neq l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{z_l\ z_k}{v_i^2\ v_j^2\ v_l\ v_k}\mu_l\mu_k +$$

$$\left.+2\sum_{\substack{i=1\\i\neq j\neq l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{z_l\ z_k}{v_i^2\ v_j\ v_l\ v_k}\mu_l\mu_k\right]\Bigg\}$$

B.3 Terceiro Momento

Para facilitar os restantes cálculos consideremos:

$$Z = \left(\sum_{i=1}^{n}\frac{y_i}{v_i}\right)^2\sum_{i=1}^{n}\frac{y_i\ z_i}{v_i} = \left(\sum_{i=1}^{n}\frac{y_i^2}{v_i^2} + \sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i\ y_j}{v_i\ v_j}\right)\times\sum_{i=1}^{n}\frac{y_i\ z_i}{v_i} =$$

$$= \sum_{i=1}^{n} \frac{y_i^3 \ z_i}{v_i^3} + 3 \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} \frac{y_i^2 \ z_i \ y_j}{v_i^2 \ v_j} + \sum_{\substack{i=1 \\ i \neq j \neq l}}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \frac{y_i \ z_i \ y_j \ y_l}{v_i \ v_j \ v_l}$$

tendo-se que:

$$\mathrm{E}\left(Z\right) = \sum_{i=1}^{n} \frac{\mu'_{3,i} \ z_i}{v_i^3} + 3 \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} \frac{\mu'_{2,i} \ z_i \ \mu_j}{v_i^2 \ v_j} + \sum_{\substack{i=1 \\ i \neq j \neq l}}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \frac{\mu_i \ z_i \ \mu_j \ \mu_l}{v_i \ v_j \ v_l}$$

$$\mathrm{E}\left(\Delta^3\right) = \sum_{l=1}^{4} T_{3,l} \tag{B.3}$$

onde

$$T_{3,1} = \mathrm{E}\left(\theta^3\right) = 0$$

$$T_{3,2} = \mathrm{E}\left(-3\frac{\tilde{\beta}_1}{d}\left(\tilde{d} - d\right)\theta^2\right) = -\frac{3\sigma^4}{d^4}\left[3a\sum_{i=1}^{n}\frac{w_i \ z_i}{v_i^3}\mu'_{i,3} + 3\sum_{i=1}^{n}\frac{u_i}{v_i^3}E\left[Z\right] + \right.$$

$$+3a\sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n}\frac{w_i \ z_j}{v_i^2 \ v_j}\mu'_{i,3} \ \mu_j + a\sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n}\frac{w_i \ z_i}{v_i^2 \ v_j}\mu'_{i,3} + a\sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n}\frac{w_i \ z_j}{v_i \ v_j^2}\mu'_{i,2} \ \mu_j -$$

$$-2a\sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n}\frac{w_i \ z_i}{v_i^3 \ v_j^2}\mu'_{i,2} \ \mu_j - 2a\sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n}\frac{w_i \ z_j}{v_i^3 \ v_j^2}\mu_i \ \mu'_{j,2} - 6a\sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n}\frac{\mu_i \ z_i}{v_i^4 \ v_j}\mu'_{i,2} \ \mu_j -$$

$$-6a\sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n}\frac{\mu_i \ z_j}{v_i^3 \ v_j^3}\mu_i \ \mu'_{j,2} - 2a\sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n}\frac{\mu_j \ z_i}{v_i^3 \ v_j^2}\mu'_{i,2} \ \mu_j - 2a\sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n}\frac{\mu_i \ z_j}{v_i^2 \ v_j^2}\mu_i \ \mu'_{j,2} +$$

$$+4a\sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n}\frac{z_i}{v_i^4 \ v_j^2}\mu'_{i,2} + 4a\sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n}\frac{z_j}{v_i^3 \ v_j^3} \ \mu'_{i,2} \ \mu_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\mu_i}{v_i^2 \ v_j}E\left[Z\right] -$$

$$-2\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{v_i^2 \ v_j^2}E\left[Z\right] + a\sum_{\substack{i=1 \\ i \neq j \neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{w_i \ z_l}{v_i \ v_j \ v_l}\mu'_{i,2} \ \mu_l -$$

$$-2a\sum_{\substack{i=1 \\ i \neq j \neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{w_i \ z_l}{v_i^2 \ v_j^2 \ v_l}\mu_i \ \mu_j \ \mu_l - 6a\sum_{\substack{i=1 \\ i \neq j \neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{\mu_i \ z_l}{v_i^3 \ v_j \ v_l}\mu_i \ \mu_j \ \mu_l -$$

$$-2a\sum_{\substack{i=1 \\ i \neq j \neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{\mu_i \ z_j}{v_i^2 \ v_j^2 \ v_l}\mu_i \ \mu_j \ \mu_l - 2a\sum_{\substack{i=1 \\ i \neq j \neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{\mu_l \ z_i}{v_i^3 \ v_j \ v_l}\mu'_{i,2} \ \mu_j -$$

$$-2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{\mu_l\ z_j}{v_i^2\ v_j^2\ v_l}\mu_i\ \mu'_{j,2} - 2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{\mu_l\ z_l}{v_i^2\ v_j\ v_l^2}\mu_i\ \mu_j\ \mu_l +$$

$$+4a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{z_l}{v_i^3\ v_j^2\ v_l}\mu'_{i,2}\ \mu_l + 4a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{z_i}{v_i^3\ v_j^2\ v_l}\mu'_{i,2}\ \mu_l +$$

$$+4a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{z_l}{v_i^2\ v_j^2\ v_l^2}\mu_i\ \mu'_{l,2} - 2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j\neq l\neq k}}^{n}\frac{\mu_l\ z_k}{v_i^2\ v_j\ v_l\ v_k}\mu_i\ \mu_j\ \mu_k +$$

$$+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j\neq l\neq k}}^{n}\frac{z_k}{v_i^2\ v_j^2\ v_l\ v_k}\mu_i\mu_l\mu_k\Bigg]$$

$$T_{3,3} = \mathrm{E}\left(3\frac{\tilde{\beta}_1^2}{d^2}\left(\tilde{d}-d\right)^2\theta\right) = \frac{3\sigma^4}{d^5}\left[12a\sum_{i=1}^{n}\frac{f_i\ z_i^2}{v_i^7}\left(1-2av_i\right)\mu'_{i,3} +\right.$$

$$+4a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i\ z_i^2}{v_i^5\ v_j^3}\mathbf{b}_{1,1}\ \mu_i + 4a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i\ z_i^2}{v_i^5\ v_j^2}\mathbf{b}_{1,2}\ \mu'_{i,3} +$$

$$+12a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i\ z_i^2}{v_i^6\ v_j^3}\mathbf{b}_{1,3}\ \mu'_{i,2}\ \mu_j + 12a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i\ z_j^2}{v_i^4\ v_j^5}\mathbf{b}_{1,4}\ \mu'_{j,3} +$$

$$+8a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i\ z_i\ z_j}{v_i^5\ v_j^3}\mathbf{b}_{1,5}\ \mu'_{i,2}\ \mu_j + 2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i\ z_i\ z_j}{v_i^5\ v_j^4}\mathbf{b}_{1,6}\ \mu_i\ \mu'_{j,2} -$$

$$-2a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_i^2}{v_i^4\ v_j^2}\mu'_{i,2}\ \mu_j - 2a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_j^2}{v_i^2\ v_j^4}\mu'_{j,3} - 4a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{z_i\ z_j}{v_i^3\ v_j^3}\mu_i\ \mu'_{j,2} +$$

$$+4a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{f_i\ z_l^2}{v_i^2\ v_j^2\ v_l^3}\mathbf{b}_{1,7}\ \mu_i\ \mu'_{l,2} + 12a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{f_i\ z_l^2}{v_i^4\ v_j^3\ v_l^2}\mathbf{b}_{1,8}\ \mu_j\ \mu'_{l,2} +$$

$$+4a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{f_i\ z_i^2}{v_i^4\ v_j^2\ v_l}\mathbf{b}_{1,9}\ \mu'_{i,2}\ \mu_l + 4a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{f_i\ z_j^2}{v_i^2\ v_j^4\ v_l}\mathbf{b}_{1,10}\ \mu'_{j,2}\ \mu_l +$$

$$+4a\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i\ z_l^3}{v_i^2\ v_j^2\ v_l^3}\mathbf{b}_{1,11}\ \mu'_{l,3}+4a\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i\ z_j\ z_l}{v_i^5\ v_j^3\ v_l^3}\mathbf{b}_{1,12}\ \mu_i\ \mu_j\ \mu_l+$$

$$+8a\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i\ z_i\ z_l}{v_i^5\ v_j^2\ v_l}\mathbf{b}_{1,13}\ \mu'_{i,2}\ \mu_l+24a\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i\ z_i\ z_l}{v_i^5\ v_j^3\ v_l}\mathbf{b}_{1,14}\ \mu_i\ \mu_j\ \mu_l+$$

$$+24a\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i\ z_j\ z_l}{v_i^4\ v_j^4\ v_l}\mathbf{b}_{1,15}\ \mu'_{j,2}\ \mu_l+8a\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i\ z_i\ z_l}{v_i^3\ v_j^2\ v_l^2}\mathbf{b}_{1,16}\ \mu_i\ \mu'_{l,2}+$$

$$+8a\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i\ z_j\ z_l}{v_i^2\ v_j^3\ v_l^2}\mathbf{b}_{1,17}\ \mu_j\ \mu'_{l,2}-2a^2\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_l^2}{v_i^2\ v_j^2\ v_l^2}\mu_j\ \mu'_{l,2}-$$

$$-9a^2\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_j\ z_l}{v_i^4\ v_j\ v_l}\mu_i\ \mu_j\ \mu_l-4a^2\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_i\ z_l}{v_i^3\ v_j^2\ v_l}\mu_i\ \mu_j\ \mu_l-$$

$$-4a^2\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{z_j\ z_l}{v_i^2\ v_j^3\ v_l}\mu'_{j,2}\ \mu_l+4a\sum_{\substack{i=1\\i\neq j\neq l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{f_i\ z_k^2}{v_i^2\ v_j^2\ v_l\ v_k^2}\mathbf{b}_{1,18}\ \mu_l\ \mu'_{k,2}+$$

$$+4a\sum_{\substack{i=1\\i\neq j\neq l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{f_i\ z_l\ z_k}{v_i^3\ v_j^2\ v_l\ v_k}\mathbf{b}_{1,19}\ \mu_i\ \mu_l\ \mu_k+$$

$$+4a\sum_{\substack{i=1\\i\neq j\neq l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{f_i\ z_l\ z_k}{v_i^4\ v_j^3\ v_l\ v_k}\mathbf{b}_{1,20}\ \mu_j\ \mu_l\ \mu_k-$$

$$-8a^2\sum_{\substack{i=1\\i\neq j\neq l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{z_l\ z_k}{v_i^2\ v_j^2\ v_l\ v_k}\mu_j\ \mu_l\ \mu_k+$$

$$+8a\sum_{\substack{i=1\\i\neq j\neq l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{f_i\ z_i\ z_k}{v_i^2\ v_j^2\ v_l\ v_k}\mathbf{b}_{1,21}\ \mu_i\ \mu_l\ \mu_k+$$

$$+8a\sum_{\substack{i=1\\i\neq j\neq l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{f_i\ z_l\ z_k}{v_i^2\ v_j^2\ v_l^2\ v_k}\mathbf{b}_{1,22}\ \mu'_{l,2}\ \mu_k+$$

$$\left.+4a\sum_{\substack{i=1\\i\neq j\neq l\neq k\neq t}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{f_i\ z_k\ z_t}{v_i^2\ v_j^2\ v_l\ v_k\ v_{gt}}\mathbf{b}_{1,23}\mu_l\mu_k\mu_t\right]$$

$$T_{3,4} = \mathrm{E}\left(-\frac{\tilde{\beta}_1^3}{d^3}\left(\tilde{d}-d\right)^3\right) = -\frac{\left(\tilde{d}-d\right)^3}{d^3}\mathrm{E}\left(\tilde{\beta}_1^3\right) =$$

$$= -\frac{\sigma^4}{d^3}\left[12a^2\sum_{i=1}^{n}\frac{f_i^2}{v_i^4}\mathbf{b}_{3,1}\mathrm{E}\left[\tilde{\beta}_1^3\right] + 12a^2\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i^2}{v_i^2\ v_j^2}\mathbf{b}_{3,2}\mathrm{E}\left[\tilde{\beta}_1^3\right] -\right.$$

$$\left. -8a^2\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{2\ f_i\ f_j\ +1}{v_i^2\ v_j^2}\mathrm{E}\left[\tilde{\beta}_1^3\right]\right] -$$

$$-\frac{\sigma^6}{d^3}\left[-15\underbrace{\sum_{i=1}^{n}\left(\frac{1-av_i}{v_i^2}\right)^3}_{\mathbf{b}_{3,3}}\mathrm{E}\left[\tilde{\beta}_1^3\right] + 3\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{1}{v_i^4\ v_j^2}\mathbf{b}_{3,4}\mathrm{E}\left[\tilde{\beta}_1^3\right] +\right.$$

$$\left. +\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{1}{v_i^2\ v_j^2\ v_l^2}\mathbf{b}_{3,5}E\left[\tilde{\beta}_1^3\right]\right]$$

com

$$\mathbf{b}_{1,1} = 3a^2v_iv_j - 6av_iv_j + 3v_j + a^2v_i^3 - av_i^3v_j - av_i^2$$

$$\mathbf{b}_{1,2} = av_iv_j - v_i - v_j$$

$$\mathbf{b}_{1,3} = -av_i^2v_j + v_i^2 - v_iv_j^2 + v_j^2$$

$$\mathbf{b}_{1,4} = -av_i^2v_j^3 + v_i^2 - av_iv_j^2 + v_j^2$$

$$\mathbf{b}_{1,5} = 3a^2v_iv_j^2 - 6av_j^2 + a^2v_i^2v_j - av_i^2 - av_iv_j + 3v_iv_j$$

$$\mathbf{b}_{1,6} = -av_i^2v_j + v_i^2 - av_iv_j^2 + v_j^2$$

$$\mathbf{b}_{1,7} = av_jv_l - v_l - v_j$$

$$\mathbf{b}_{1,8} = av_i^2v_j + v_i^3 - a^2v_iv_j^2 + v_j^2$$

$$\mathbf{b}_{1,9} = -av_j + 3$$

$$\mathbf{b}_{1,10} = -av_j + 3$$

$$\mathbf{b}_{1,11} = -av_j + 3v_l^2$$

$$\mathbf{b}_{1,12} = 3a^2v_i^2v_j^2v_l^2 - 6av_iv_j^2v_l^2 - 3v_j^2v_l^2 + 2a^2v_i^3v_jv_l^2 - 2av_i^3v_l^2 - 2av_i^2v_jv_l^2 +$$

$$+6v_i^2v_jv_l^2 - 2av_i^2v_j^2v_l + 6v_i^2v_j^2$$

$$\mathbf{b}_{1,13} = a^2v_j - av_i^2 - av_iv_j + 3v_iv_j$$

$$\mathbf{b}_{1,14} = -av_i^2v_j + v_i^2 - av_iv_j^2 + v_j^2$$

$$\mathbf{b}_{1,15} = -av_i^2v_j + v_i^2 - av_iv_j^2 + v_j^2$$

$$\mathbf{b}_{1,16} = -av_j + 3$$

$$\mathbf{b}_{1,17} = -av_j + 3$$

$$\mathbf{b}_{1,18} = -av_j + 3$$

$$\mathbf{b}_{1,19} = a^2v_iv_j - av_i - av_j + 3v_j$$

$$\mathbf{b}_{1,20} = -3av_i^2v_jv_l^2 + 3v_i^2v_l^2 - 3av_iv_j^2v_l^2 + 3v_j^2v_l^2 - 2av_i^2v_j^2v_l + 6v_i^2v_j^2$$

$$\mathbf{b}_{1,21} = -av_j + 3$$

$$\mathbf{b}_{1,22} = -av_j + 3$$

$$\mathbf{b}_{1,23} = -av_j + 3$$

$$\mathbf{b}_{3,1} = 2av_i - 3 + av_i^2$$

$$\mathbf{b}_{3,2} = av_j - 1$$

$$\mathbf{b}_{3,3} = \left(\frac{1-av_i}{v_i^2}\right)^3$$

$$\mathbf{b}_{3,4} = 3a^3v_i^2v_j - 3a^2v_i^2 - 6a^2v_iv_j + (18a-2)\,v_iv_j + (3a-5)\,v_j - 8$$

$$\mathbf{b}_{3,5} = a^3v_iv_jv_l - (3a^2-2a)\,v_jv_l + 5av_l + 2a - 15$$

B.4 Quarto Momento

$$\Delta^4 = (\theta-\chi)^4 = \underbrace{\theta^4}_{I} - \underbrace{4\theta^3\chi}_{II} + \underbrace{6\theta^2\chi^2}_{III} - \underbrace{4\theta\chi^3}_{IV} + \underbrace{\chi^4}_{V}$$

com

$$\begin{cases} \chi = \frac{\tilde{\beta}_1}{d}\left(\tilde{d} - d\right) \\ \text{e} \\ \theta = \frac{1}{d}\left[a\sum_{i=1}^{n}\frac{\varepsilon_i y_i}{v_i} - \sum_{i=1}^{n}\frac{\varepsilon_i y_i}{v_i^2} - \sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{\varepsilon_i y_j}{v_i v_j}\right] \end{cases}$$

Queremos calcular $\mathrm{E}\left(\Delta^4\right)$. Comecemos por estudar o termo I, tendo-se:

$$I/1 \rightarrow \frac{a}{d}\left(\sum_{i=1}^{n}\frac{\varepsilon_i y_i}{v_i}\right)\theta^3$$

e

$$I/2 \rightarrow -\frac{1}{d}\left(\sum_{i=1}^{n}\frac{\varepsilon_i y_i}{v_i^2}\right)\theta^3$$

Nota: os termos destes dois desenvolvimentos correspondem-se, no primeiro aparece o factor a e no segundo o factor $-\frac{1}{v_i}$. Um termo tem valor médio $\neq 0$ se e só se o mesmo acontece com o termo correspondente, fazendo-se a substituição de a por $-\sum_{i=1}^{n}\frac{1}{v_i}$ para o cálculo do valor médio.

e

$$I/3 \rightarrow -\frac{1}{d}\left(\sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{\varepsilon_i y_j}{v_i v_j}\right)\theta^3$$

Vamos agora calcular o valor médio de cada um destes subtermos.

$$\mathrm{E}\left(I/1\right) = \frac{a}{d^4}\sigma^4\left\{a^3\left[3\sum_{i=1}^{n}\frac{y_i^4}{v_i^4} + 3\sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{y_i^2 y_j^2}{v_i^2 v_j^3}\right] - a^2\left[9\sum_{i=1}^{n}\frac{y_i^4}{v_i^5} + 9\sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{y_i^2 y_j^2}{v_i^2 v_j^3}\right.\right.$$

$$\left. + 9\sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{y_i^3 y_j}{v_i^4 v_j} + 9\sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{y_i y_j^3}{v_i^2 v_j^3} + 9\sum_{\substack{i=1j=1l=1 \\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_i^2 y_j y_l}{v_i^2 v_j^2 v_l}\right] + a\left[9\sum_{i=1}^{n}\frac{y_i^4}{v_i^6} + 2\sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{y_i y_j^3}{v_i^2 v_j^3} + \right.$$

$$+12\sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{y_i^2 y_j^2}{v_i^2 v_j^4} + 12\sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{y_i^2 y_j^2}{v_i^3 v_j^3} + 18\sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{y_i^3 y_j}{v_i^5 v_j} + 6\sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{y_i y_j^3}{v_i^3 v_j^3} + \sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{y_i y_j^3}{v_i v_j^4} +$$

$$+9\sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{y_i y_j^3}{v_i^2 v_j^4} + 3\sum_{\substack{i=1j=1 \\ i\neq j}}^{n}\sum^{n}\frac{y_j^4}{v_i^2 v_j^4} + 6\sum_{\substack{i=1j=1l=1 \\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_i^2 y_j y_l}{v_i^2 v_j^3 v_l} + 24\sum_{\substack{i=1j=1l=1 \\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_i^2 y_j y_l}{v_i^3 v_j^2 v_l}$$

$$\left.\left. + 3\sum_{\substack{i=1j=1l=1 \\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_i^2 y_j^2}{v_i^2 v_j^2 v_l^2} + 9\sum_{\substack{i=1j=1l=1 \\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_i^2 y_j y_l}{v_i^4 v_j v_l} + 6\sum_{\substack{i=1j=1l=1 \\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_j^3 y_l}{v_i^2 v_j^3 v_l} + 5\sum_{\substack{i=1j=1l=1 \\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_i y_j^2 y_l}{v_i^2 v_j^2 v_l^2} + \right.\right.$$

$$+\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 y_l}{v_i^2 v_j^2 v_l}+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i^2 y_j y_k}{v_i^2\ v_j^2\ v_l\ v_k}+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_j y_l y_k}{v_i\ v_j\ v_l^2\ v_k}+$$
$$i\neq j\neq l \qquad i\neq j\neq l\neq k \qquad i\neq j\neq l\neq k$$

$$+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_j y_l y_k}{v_i^2\ v_j\ v_l^2\ v_k}\Bigg]-\Bigg[3\sum_{\substack{i=1}}^{n}\frac{y_i^4}{v_i^7}+\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i^2 y_j^2}{v_i^3 v_j^3}+\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i y_j^3}{v_i^3 v_j^3}+\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i y_j^3}{v_i^2 v_j^4}+$$
$$i\neq j \qquad i\neq j \qquad i\neq j$$

$$+\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_j^4}{v_i^2 v_j^4}+8\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i^2 y_j^2}{v_i^3 v_j^4}+9\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i^3 y_j}{v_i^6 v_j}+8\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i y_j^3}{v_i^3 v_j^4}+2\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i y_j^3}{v_i^2 v_j^5}+$$
$$i\neq j \qquad i\neq j \qquad i\neq j \qquad i\neq j \qquad i\neq j$$

$$+9\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i^2 y_j^2}{v_i^5 v_j^2}+2\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_j^4}{v_i^2 v_j^5}+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i y_j^3}{v_i^4 v_j^3}+19\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j y_l}{v_i^3 v_j^2 v_l}+$$
$$i\neq j \qquad i\neq j \qquad i\neq j \qquad i\neq j\neq l$$

$$+\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j^3 y_l}{v_i^2 v_j^3 v_l}+8\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j y_l}{v_i^4 v_j^2 v_l}+9\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j^2}{v_i^3 v_j^2 v_l^2}+9\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j y_l}{v_i^5 v_j v_l}+$$
$$i\neq j\neq l \qquad i\neq j\neq l \qquad i\neq j\neq l \qquad i\neq j\neq l$$

$$+8\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j^3 y_l}{v_i^2 v_j^4 v_l}+8\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 y_l}{v_i^3 v_j^2 v_l^2}+\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 y_l}{v_i^2 v_j^2 v_l^2}+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j^3 y_l}{v_i^2 v_j^3 v_l^2}+$$
$$i\neq j\neq l \qquad i\neq j\neq l \qquad i\neq j\neq l \qquad i\neq j\neq l$$

$$+9\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 y_l}{v_i^4 v_j^2 v_l}+9\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i^2 y_l y_k}{v_i^3\ v_j^2\ v_l\ v_k}+9\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_j y_l y_k}{v_i^3\ v_j\ v_l^2\ v_k}+$$
$$i\neq j\neq l \qquad i\neq j\neq l\neq k \qquad i\neq j\neq l\neq k$$

$$+9\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_j^2 y_l y_k}{v_i^2\ v_j^2\ v_l^2\ v_k}+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_j y_l y_k}{v_i^4\ v_j\ v_l\ v_k}+$$
$$i\neq j\neq l\neq k \qquad i\neq j\neq l\neq k$$

$$+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{y_j\ y_l\ y_k\ y_t}{v_i^2\ v_j\ v_l\ v_k\ v_t}\Bigg]\Bigg\}$$
$$i\neq j\neq l\neq k\neq t$$

Relativamente a $\mathrm{E}\,(I/2)$ já referimos que basta substituir em $\mathrm{E}\,(I/1)$ o factor a pelo factor $-\sum_{i=1}^{n}\frac{1}{v_i}$.

$$\mathrm{E}\,(I/3)=-\tfrac{1}{d^3}\sigma^4\Bigg\{a^3\Bigg[3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i^3 y_j}{v_i^4 v_j}+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j y_l}{v_i^2 v_j^2 v_l}\Bigg]-a^2\Bigg[2\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^3 y_j}{v_i^5 v_j}$$
$$i\neq j \qquad i\neq j,i\neq l$$

$$+12\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 y_l}{v_i^2 v_j^3 v_l}+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 y_l}{v_i^3 v_j^2 v_l}+9\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j y_l}{v_i^4 v_j v_l}+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j^3 y_l}{v_i^2 v_j^3 v_l}+$$
$$i\neq j,i\neq l \qquad i\neq j,i\neq l \qquad i\neq j,i\neq l$$

$$+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i^2 y_l y_k}{v_i^2\ v_j^2\ v_l\ v_k}+6\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_j y_l y_k}{v_i^2\ v_j^2\ v_l\ v_k}\Bigg]+a\Bigg[9\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i^3 y_j}{v_i^6 v_j}+$$

$$+\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 y_l}{v_i^2 v_j^3 v_l}+\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j^3 y_l}{v_i^2 v_j^3 v_l}+7\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 y_l}{v_i^2 v_j^4 v_l}+18\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j y_l}{v_i^3 v_j^3 v_l}+$$

$$+18\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j y_l}{v_i^5 v_j v_l}+\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j y_l}{v_i^4 v_j v_l}+8\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j^3 y_l}{v_i^2 v_j^4 v_l}+9\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 y_l}{v_i^4 v_j^2 v_l}+$$

$$+18\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_j y_l y_k}{v_i^2\ v_j^3\ v_l\ v_k}+18\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_j^2 y_l y_k}{v_i^2\ v_j^3\ v_l\ v_k}+9\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_l^2 y_k}{v_i^2\ v_j^2\ v_l^2\ v_k}+$$

$$+9\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_j y_l y_k}{v_i^4\ v_j\ v_l\ v_k}+\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_j^2 y_l y_k}{v_i^2\ v_j^2\ v_l\ v_k}+6\sum_{\substack{i=1\\ i\neq j,l,k;j\neq l,k;l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{y_i y_l\ y_k\ y_t}{v_i^2\ v_j^2\ v_l\ v_k\ v_t}+$$

$$+3\sum_{\substack{i=1\\ i\neq j,l,k;j\neq l,k;l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{y_j y_l\ y_k\ y_t}{v_i^2\ v_j\ v_l\ v_k\ v_t}\Bigg]-\Bigg[3\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i^3 y_j}{v_i^7 v_j}+2\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j y_l}{v_i^3 v_j^3 v_l}$$

$$+2\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j^3 y_l}{v_i^2 v_j^4 v_l}+16\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 y_l}{v_i^3 v_j^4 v_l}+9\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j y_l}{v_i^6 v_j v_l}+2\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j^3 y_l}{v_i^2 v_j^5 v_l}+$$

$$+9\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 y_l}{v_i^5 v_j^2 v_l}+2\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^4 y_j}{v_i^5 v_j^2 v_l}+3\sum_{\substack{i=1\\ i\neq j,i\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j^3 y_l}{v_i^4 v_j^3 v_l}+$$

$$+2\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_j^2 y_l y_k}{v_i^2\ v_j^3\ v_l\ v_k}+18\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_j y_l y_k}{v_i^3\ v_j^3\ v_l\ v_k}+16\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_j^2 y_l y_k}{v_i^2\ v_j^4\ v_l\ v_k}+$$

$$+17\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_l^2 y_k}{v_i^3\ v_j^2\ v_l^2\ v_k}+9\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_j y_l y_k}{v_i^5\ v_j\ v_l\ v_k}+\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_j^2 y_l y_k}{v_i^2\ v_j^2\ v_l^2\ v_k}+$$

$$+3\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_l^3 y_k}{v_i^2\ v_j^2\ v_l^3\ v_k}+9\sum_{\substack{i=1\\ i\neq j,l;j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_j^2 y_l y_k}{v_i^4\ v_j^2\ v_l\ v_k}+18\sum_{\substack{i=1\\ i\neq j,l,k;j\neq l,k;l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{y_i y_l\ y_k\ y_t}{v_i^3\ v_j^2\ v_l\ v_k\ v_t}+$$

$$+9\sum_{\substack{i=1\\ i\neq j,l,k;j\neq l,k;l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{y_l^3\ y_k\ y_t}{v_i^2\ v_j^2\ v_l^2\ v_k\ v_t}+3\sum_{\substack{i=1\\ i\neq j,l,k;j\neq l,k;l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{y_j y_l\ y_k\ y_t}{v_i^4\ v_j\ v_l\ v_k\ v_t}+$$

$$+3\sum_{\substack{i=1\\ i\neq j,l,k,t;j\neq l,k,t;l\neq k,t;k\neq t}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\sum_{u=1}^{n}\frac{y_l\ y_k\ y_t\ y_u}{v_i^2\ v_j^2\ v_l\ v_k\ v_t\ v_u}\Bigg]\Bigg\}$$

Relembremos que

$$\begin{cases} \chi = \frac{\tilde{\beta}_1}{d}\left(\tilde{d}-d\right) \\ \\ \left(\tilde{d}-d\right) = 2a\sum_{i=1}^{n}\frac{f_i\varepsilon_i}{v_i} + a\sum_{i=1}^{n}\frac{\varepsilon_i^2}{v_i} - \sum_{i=1}^{n}\frac{\varepsilon_i^2}{v_i^2} - \sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{\varepsilon_i\varepsilon_j}{v_iv_j} \\ \\ \tilde{\beta}_1 = \frac{1}{d}\sum_{i=1}^{n}\frac{y_iz_i}{v_i} \quad \text{com} \quad z_i = af_i - b \end{cases}$$

tendo-se deste modo

$$\chi = \frac{1}{d^2}\left[2a\sum_{i=1}^{n}\frac{y_i\varepsilon_if_iz_i}{v_i^2} + 2a\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{\varepsilon_if_i\varepsilon_jf_j}{v_iv_j} + a\sum_{i=1}^{n}\frac{y_i\varepsilon_i^2z_i}{v_i^2} + a\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{\varepsilon_i^2y_jz_j}{v_iv_j} - \right.$$

$$\left. -\sum_{i=1}^{n}\frac{y_i\varepsilon_i^2z_i}{v_i^3} - \sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{\varepsilon_i^2y_jz_j}{v_i^2v_j} - 2\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i\varepsilon_iz_i\varepsilon_j}{v_i^2v_j} - \sum_{\substack{i=1 \\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{\varepsilon_i\varepsilon_jy_lz_l}{v_iv_jv_l}\right]$$

Vamos agora estudar o termo II, seguindo o mesmo raciocínio obtemos

$$II/1 \rightarrow -\frac{8a}{d^2}\left(\sum_{i=1}^{n}\frac{y_i\varepsilon_if_iz_i}{v_i^2}\right)\theta^3$$

$$II/2 \rightarrow -\frac{8a}{d^2}\left(\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{\varepsilon_if_i\varepsilon_jf_j}{v_iv_j}\right)\theta^3$$

$$II/3 \rightarrow -\frac{4a}{d^2}\left(\sum_{i=1}^{n}\frac{y_i\varepsilon_i^2z_i}{v_i^2}\right)\theta^3$$

$$II/4 \rightarrow -\frac{4a}{d^2}\left(\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{\varepsilon_i^2y_jz_j}{v_iv_j}\right)\theta^3$$

$$II/5 \rightarrow \frac{4}{d^2}\left(\sum_{i=1}^{n}\frac{y_i\varepsilon_i^2z_i}{v_i^3}\right)\theta^3$$

$$II/6 \rightarrow \frac{4}{d^2}\left(\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{\varepsilon_i^2y_jz_j}{v_i^2v_j}\right)\theta^3$$

$$II/7 \to \frac{8}{d^2}\left(\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i\varepsilon_i z_i\varepsilon_j}{v_i^2 v_j}\right)\theta^3$$

$$II/8 \to \frac{4}{d^2}\left(\sum_{\substack{i=1 j=1 l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{\varepsilon_i\varepsilon_j y_l z_l}{v_i v_j v_l}\right)\theta^3$$

Verificamos que

$$\begin{cases} \mathrm{E}\,(II/1) \approx \mathrm{E}\,(II/2) \approx \mathrm{E}\,(I/1) \approx \mathrm{E}\,(I/2) \\ \\ \mathrm{E}\,(II/3)\,,\ \mathrm{E}\,(II/4)\,,\ E\,(II/5) \text{ e } \mathrm{E}\,(II/6) \quad \text{são nulos} \\ \\ \mathrm{E}\,(II/7) \approx \mathrm{E}\,(II/8) \end{cases}$$

Nota: os termos não nulos de $\mathrm{E}\,(II/1)$ e de $\mathrm{E}\,(II/2)$ são os correspondentes aos termos não nulos de $\mathrm{E}\,(I/1)$ e de $\mathrm{E}\,(I/2)$, visto que se tem a mesma situação de emparelhamento. As sub-parcelas correspondentes à família do meio são todas nulas porque nunca se consegue efectuar o emparelhamento.

$$\mathrm{E}\,(II/1) = -\frac{8a}{d^5}\sigma^4\left\{a^3\left[3\sum_{i=1}^{n}\frac{y_i^2 f_i z_i}{v_i^5} + 3\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i^2 y_j^2 f_j z_j}{v_i^2 v_j^3}\right] - a^2\left[9\sum_{i=1}^{n}\frac{y_i^4 f_i z_i}{v_i^6}\right.\right.$$

$$+6\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i^2 f_i z_i y_j}{v_i^3 v_j^3} + 9\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i^2 f_i z_i y_j^2}{v_i^4 v_j^2} + 9\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i^3 f_i z_i y_j}{v_i^5 v_j} + 3\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i f_i z_i y_j^3}{v_i^3 v_j^3} +$$

$$\left.+6\sum_{\substack{i=1 j=1 l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_i^2 f_i z_i y_j y_l}{v_i^3 v_j^2 v_l} + 3\sum_{\substack{i=1 j=1 l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_i^2 y_j f_j z_j y_l}{v_i^2 v_j^3 v_l}\right] + a\left[9\sum_{i=1}^{n}\frac{y_i^4 f_i z_i}{v_i^7} +\right.$$

$$+\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i^2 f_i z_i y_j^2}{v_i^4 v_j^2} + \sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i f_i z_i y_j^3}{v_i^3 v_j^3} + 3\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i^2 f_i z_i y_j^2}{v_i^3 v_j^4} + 12\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i^2 y_j^2 f_j z_j}{v_i^3 v_j^4} +$$

$$+18\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i^3 f_i z_i y_j}{v_i^6 v_j} + 6\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i y_j^3 f_j z_j}{v_i^3 v_j^4} + \sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i y_j^3 f_j z_j}{v_i v_j^5} + 5\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i f_i z_i y_j^3}{v_i^3 v_j^4} +$$

$$+9\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i^2 f_i z_i y_j^2}{v_i^5 v_j^2} + 3\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_j^4 f_j z_j}{v_i^2 v_j^5} + 4\sum_{\substack{i=1 j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i y_j^3 f_j z_j}{v_i^2 v_j^5} + 6\sum_{\substack{i=1 j=1 l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_i^2 f_i z_i y_j y_l}{v_i^3 v_j^3 v_l} +$$

$$+12\sum_{\substack{i=1 j=1 l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_i^2 y_j f_j z_j y_l}{v_i^3 v_j^3 v_l} + 12\sum_{\substack{i=1 j=1 l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_i^2 f_i z_i y_j y_l}{v_i^4 v_j^2 v_l} + 3\sum_{\substack{i=1 j=1 l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{y_i^2 f_i z_i y_l^2}{v_i^3 v_j^2 v_l^2} +$$

$$+9\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j\neq l}}^{n}\sum_{l=1}^{n}\frac{y_i^2 f_i z_i y_j y_l}{v_i^5 v_j v_l}+6\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j\neq l}}^{n}\sum_{l=1}^{n}\frac{y_j^3 f_j z_j y_l}{v_i^2 v_j^4 v_l}+5\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{y_i y_j^2 y_l f_l z_l}{v_i^2 v_j^2 v_l^3}+$$

$$+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j\neq l}}^{n}\sum_{l=1}^{n}\frac{y_i f_i z_i y_j^2 y_l}{v_i^3 v_j^2 v_l}+3\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j\neq l\neq k}}^{n}\frac{y_i^2 f_i z_i y_l y_k}{v_i^3\ v_j^2\ v_l\ v_k}+3\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j\neq l\neq k}}^{n}\frac{y_i y_j y_l f_l z_l y_k}{v_i\ v_j\ v_l^2\ v_k}+$$

$$+3\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j\neq l\neq k}}^{n}\frac{y_i f_i z_i y_j y_l y_k}{v_i^3\ v_j\ v_l^2\ v_k}\Bigg]-\Bigg[3\sum_{i=1}^{n}\frac{y_i^4 f_i z_i}{v_i^8}+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i^2 y_j^2 f_j z_j}{v_i^3 v_j^4}+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i y_j^3 f_j z_j}{v_i^3 v_j^4}+$$

$$+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i f_i z_i y_j^3}{v_i^3 v_j^4}+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_j^4 f_i z_j}{v_i^2 v_j^5}+8\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i^2 f_i z_i y_j^2}{v_i^4 v_j^4}+9\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i^3 f_i z_i y_j}{v_i^7 v_j}+$$

$$+8\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i y_j^3 f_j z_j}{v_i^3 v_j^5}+2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i f_i z_i y_j^3}{v_i^3 v_j^5}+9\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i^2 f_i z_i y_j^2}{v_i^6 v_j^2}+2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_j^4 f_j z_j}{v_i^2 v_j^6}+$$

$$+3\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i f_i z_i y_j^3}{v_i^5 v_j^3}+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{y_i^2 y_j f_j z_j y_l}{v_i^3 v_j^3 v_l}+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{y_j^3 f_j z_j y_l}{v_i^2 v_j^4 v_l}+$$

$$+18\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{y_i^2 f_i z_i y_j y_l}{v_i^4 v_j^3 v_l}+8\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{y_i^2 y_j f_j z_j y_l}{v_i^4 v_j^3 v_l}+9\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{y_i^2 f_i z_i y_l^2}{v_i^4 v_j^2 v_l^2}+$$

$$+9\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{y_i^2 f_i z_i y_j y_l}{v_i^6 v_j v_l}+8\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{y_j f_j z_j y_l}{v_i^2 v_j^5 v_l}+8\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{y_i y_j^2 y_l f_l z_l}{v_i^3 v_j^2 v_l^3}+$$

$$+\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{y_i f_i z_i y_j^2 y_l}{v_i^3 v_j^2 v_l^2}+3\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{y_j^3 y_l f_l z_l}{v_i^2 v_j^3 v_l^3}+9\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{y_i f_i z_i y_j^2 y_l}{v_i^5 v_j^2 v_l}+$$

$$+9\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j\neq l\neq k}}^{n}\frac{y_i^2 f_i z_i y_l y_k}{v_i^4\ v_j^2\ v_l\ v_k}+9\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j\neq l\neq k}}^{n}\frac{y_i y_j y_l f_l z_l y_k}{v_i^3\ v_j\ v_l^3\ v_k}+$$

$$+9\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j\neq l\neq k}}^{n}\frac{y_j^2 y_l f_l z_l y_k}{v_i^2\ v_j^2\ v_l^3\ v_k}+3\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j\neq l\neq k}}^{n}\frac{y_i f_i z_i y_j y_l y_k}{v_i^5\ v_j\ v_l\ v_k}+$$

$$+3\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1}}^{n}\sum_{\substack{t=1\\ i\neq j\neq l\neq k\neq t}}^{n}\frac{y_j y_l y_k f_k z_k y_t}{v_i^2\ v_j\ v_l\ v_k^2\ v_t}\Bigg]\Bigg\}$$

$$\mathrm{E}\,(II/2)=-\frac{8a}{d^5}\sigma^4\left\{a^3\left[3\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i^3 f_i y_j z_j}{v_i^4 v_j}+3\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_i^2 y_j f_j y_l z_l}{v_i^2 v_j^2 v_l}\right]+\right.$$

$$-a^2\left[9\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i^3 f_i y_j z_j}{v_i^5 v_j}+6\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i f_i y_j^2 y_l z_l}{v_i^2 v_j^3 v_l}+3\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i f_i y_j^2 y_l z_l}{v_i^3 v_j^2 v_l}+\right.$$

$$+9\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i f_i y_j y_l z_l}{v_i^4 v_j v_l}+6\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 f_j y_l z_l}{v_i^2 v_j^3 v_l}+3\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i y_j^3 y_l z_l}{v_i^2 v_j^3 v_l}+$$

$$\left.+6\sum_{\substack{i=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i f_i y_j y_l y_k z_k}{v_i^2\ v_j^2\ v_l\ v_k}+3\sum_{\substack{i=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i^2 f_j y_l y_k z_k}{v_i^2\ v_j^2\ v_l\ v_k}\right]+$$

$$+a\left[9\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i^3 f_i y_j z_j}{v_i^6 v_j}+\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 f_j y_l z_l}{v_i^2 v_j^3 v_l}+\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i y_j^3 y_l z_l}{v_i^2 v_j^3 v_l}+\right.$$

$$+3\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i f_i y_j^2 y_l z_l}{v_i^2 v_j^4 v_l}+12\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j f_j y_l z_l}{v_i^3 v_j^2 v_l}+18\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i^2 f_i y_j y_l z_l}{v_i^5 v_j v_l}+$$

$$+6\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 f_j y_l z_l}{v_i^3 v_j^3 v_l}+\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 f_j y_l z_l}{v_i v_j^4 v_l}+5\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i y_j^3 y_l z_l}{v_i^2 v_j^4 v_l}+$$

$$+9\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i f_i y_j^2 y_l z_l}{v_i^4 v_j^2 v_l}+3\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j^3 f_j y_l z_l}{v_i^2 v_j^4 v_l}+4\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 f_j y_l z_l}{v_i^2 v_j^4 v_l}+$$

$$+6\sum_{\substack{i=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i f_i y_j y_l y_k z_k}{v_i^2\ v_j^3\ v_l\ v_k}+12\sum_{\substack{i=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i^2 f_j y_l y_k z_k}{v_i^3\ v_j^2\ v_l\ v_k}+$$

$$+12\sum_{\substack{i=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i f_i y_j y_l y_k z_k}{v_i^3\ v_j^2\ v_l\ v_k}+3\sum_{\substack{i=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i f_i y_l^2 y_k z_k}{v_i^2\ v_j^2\ v_l^2\ v_k}+$$

$$+9\sum_{\substack{i=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i f_i y_j y_l y_k z_k}{v_i^4\ v_j\ v_l\ v_k}+6\sum_{\substack{i=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_j^2 f_j y_l y_k z_k}{v_i^2\ v_j^3\ v_l\ v_k}+$$

$$+5\sum_{\substack{i=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_j^2 f_l y_k z_k}{v_i^2\ v_j^2\ v_l^2\ v_k}+\sum_{\substack{i=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{f_i y_j^2 y_l y_k z_k}{v_i^2\ v_j^2\ v_l\ v_k}+$$

$$+3\sum_{\substack{i=1\\ i\neq j,l,k,t\ ;\ j\neq l,k\ ;\ l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{y_i f_i y_l y_k y_t z_t}{v_i^2\ v_j^2\ v_l\ v_k\ v_t}+3\sum_{\substack{i=1\\ i\neq j,l,k,t\ ;\ j\neq l,k\ ;\ l\neq k}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{y_i y_j f_l y_k y_t z_t}{v_i\ v_j\ v_l^2\ v_k\ v_t}+$$

$$+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{f_i y_j y_l y_k y_t z_t}{v_i^2\ v_j\ v_l^2\ v_k\ v_t}\Bigg]-$$
$$i\neq j,l,k,t\ ;\ j\neq l,k\ ;\ l\neq k$$

$$-\Bigg[3\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i^3 f_i y_j z_j}{v_i^7 v_j}+\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l}}^{n}\sum_{l=1}^{n}\frac{y_i^2 y_j f_j y_l z_l}{v_i^3 v_j^2 v_l}+\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l}}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 f_j y_l z_l}{v_i^3 v_j^3 v_l}+$$

$$+\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l}}^{n}\sum_{l=1}^{n}\frac{f_i y_j^3 y_l z_l}{v_i^2 v_j^4 v_l}+\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l}}^{n}\sum_{l=1}^{n}\frac{y_j^3 f_j y_l z_l}{v_i^2 v_j^4 v_l}+8\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l}}^{n}\sum_{l=1}^{n}\frac{y_i f_i y_j^2 y_l z_l}{v_i^3 v_j^4 v_l}+$$

$$+9\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l}}^{n}\sum_{l=1}^{n}\frac{y_i^2 f_i y_j y_l z_l}{v_i^6 v_j v_l}+8\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l}}^{n}\sum_{l=1}^{n}\frac{y_i y_j^2 f_j y_l z_l}{v_i^3 v_j^4 v_l}+2\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l}}^{n}\sum_{l=1}^{n}\frac{f_i y_j^3 y_l z_l}{v_i^2 v_j^5 v_l}+$$

$$+9\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l}}^{n}\sum_{l=1}^{n}\frac{y_i f_i y_j^2 y_l z_l}{v_i^5 v_j^2 v_l}+2\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l}}^{n}\sum_{l=1}^{n}\frac{y_j^2 f_j y_l z_l}{v_i^2 v_j^5 v_l}+3\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l}}^{n}\sum_{l=1}^{n}\frac{f_i y_j^3 y_l z_l}{v_i^4 v_j^3 v_l}+$$

$$+\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i^2 f_j y_l y_k z_k}{v_i^3\ v_j^2\ v_l\ v_k}+\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_j^2 f_j y_l y_k z_k}{v_i^2\ v_j^3\ v_l\ v_k}+$$

$$+18\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i f_i y_j y_l y_k z_k}{v_i^3\ v_j^3\ v_l\ v_k}+8\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i^2 f_j y_l y_k z_k}{v_i^4\ v_j^2\ v_l\ v_k}+$$

$$+9\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i f_i y_l^2 y_k z_k}{v_i^3\ v_j^2\ v_l^2\ v_k}+9\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i f_i y_j y_l y_k z_k}{v_i^5\ v_j\ v_l\ v_k}+$$

$$+8\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_j^2 f_j y_l y_k z_k}{v_i^2\ v_j^4\ v_l\ v_k}+8\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_i y_j^2 f_l y_k z_k}{v_i^3\ v_j^2\ v_l^2\ v_k}+$$

$$+\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{f_i y_j^2 y_l y_k z_k}{v_i^2\ v_j^2\ v_l^2\ v_k}+3\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_j^3 f_l y_k z_k}{v_i^2\ v_j^3\ v_l^2\ v_k}+$$

$$+9\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{f_i y_j^2 y_l y_k z_k}{v_i^4\ v_j^2\ v_l\ v_k}+9\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k,t\ ;\ j\neq l,k\ ;\ l\neq k}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{y_i f_i y_l y_k y_t z_t}{v_i^3\ v_j^2\ v_l\ v_k\ v_t}+$$

$$+9\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k,t\ ;\ j\neq l,k\ ;\ l\neq k}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{y_i y_j f_l y_k y_t z_t}{v_i^3\ v_j\ v_l^2\ v_k\ v_t}+9\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k,t\ ;\ j\neq l,k\ ;\ l\neq k}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{y_j^2 f_l y_k y_t z_t}{v_i^2\ v_j^2\ v_l^2\ v_k\ v_t}+$$

$$+3\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k,t\ ;\ j\neq l,k\ ;\ l\neq k}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{f_i y_j y_l y_k y_t z_t}{v_i^4\ v_j\ v_l\ v_k\ v_t}+3\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j,l,k,t,u\ ;\ j\neq l,k,t\ ;\ l\neq k,t}}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\sum_{u=1}^{n}\frac{y_j y_l f_k y_t y_u z_u}{v_i^2\ v_j\ v_l\ v_k^2\ v_t\ v_u}\Bigg]\Bigg\}$$

Prova-se que a soma dos expoentes dos "ε"nos termos de θ^3 é sempre igual a 3. As partições de 3 são

$$\begin{cases} 3=3 \\ 3=2+1 \\ 3=1+1+1 \end{cases}$$

No primeiro caso os produtos são da forma $\varepsilon_i^4\varepsilon_j$ ou $\varepsilon_i^3\varepsilon_j\varepsilon_l$, no segundo $\varepsilon_i^3\varepsilon_j^2$; $\varepsilon_i^3\varepsilon_j\varepsilon_l$; $\varepsilon_i^2\varepsilon_j^2\varepsilon_l$ ou $\varepsilon_i^2\varepsilon_j\varepsilon_k\varepsilon_l$, e no terceiro $\varepsilon_i^2\varepsilon_j^2\varepsilon_l$; $\varepsilon_i^2\varepsilon_j\varepsilon_k\varepsilon_l$ ou $\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l\varepsilon_t$. Em todos estes casos o valor médio é nulo vindo

$$\mathrm{E}\,(II/7)=\mathrm{E}\,(II/8)=0.$$

Passemos agora ao termo $III \to 6\theta^2\chi^2$.

Comecemos por recordar que

$$\theta=\frac{1}{d}\left[a\sum_{i=1}^{n}\frac{\varepsilon_i y_i}{v_i}-\sum_{i=1}^{n}\frac{\varepsilon_i y_i}{v_i^2}-\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{\varepsilon_i y_j}{v_i v_j}\right]$$

pelo que os termos de $\tilde{\beta}_1^2\theta\left(\tilde{d}-d\right)^2$ que multiplicado por θ dão valores médios não nulos são aqueles com uma e uma só potência ímpar de ε.

Temos também que

$$\chi^2=\tfrac{1}{d^2}\left(\sum_{i=1}^{n}\tfrac{y_i^2z_i^2}{v_i}+\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\tfrac{y_iz_iy_jz_j}{v_iv_j}\right)\left(\tilde{d}-d\right)^2$$

pelo que o termo III se resume a

$$III\to 6\frac{\tilde{\beta}_1^2}{d^2}\left(\tilde{d}-d\right)^2\to \mathrm{E}\,(III)=6\frac{\mathrm{E}\left(\tilde{\beta}_1^2\right)}{d^4}\sum_{i=1}^{16}\sum_{j=1}^{16}E_{i,j}$$

$\sum_{i=1}^{16}\sum_{j=1}^{16}E_{i,j}$ resume-se à soma dos seguintes termos:

1)

$$12a^2\sigma^4a^2\left[\sum_{i=1}^{n}\frac{f_i^2y_i^2}{v_i^4}-2a\sum_{i=1}^{n}\frac{y_i^2f_i^2}{v_i^5}-2a\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_if_i^2y_j}{v_i^4v_j}+\sum_{i=1}^{n}\frac{y_i^2f_i^2}{v_i^6}+2\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_if_i^2y_j}{v_i^5v_j}+\right.$$

$$\left.+\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i^2y_j^2}{v_i^4v_j^2}+\sum_{\substack{i=1\\ i\neq j,l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i^2y_jy_l}{v_i^4v_jv_l}\right]$$

2)

$$8a^2\sigma^4\left[a^2\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_if_iy_jf_j}{v_i^2v_j^2}-2a\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_if_iy_jf_j}{v_i^3v_j}+(2-2a)\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_iy_j^2f_j}{v_i^2v_j^3}\right.$$
$$-2a\sum_{\substack{i=1j=1l=1\\ i\neq j,l}}^{n}\sum^{n}\sum^{n}\frac{y_if_if_jy_l}{v_i^2v_j^2v_l}+2\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_if_iy_jf_j}{v_i^3v_j^3}+4\sum_{\substack{i=1j=1l=1\\ i\neq j,l}}^{n}\sum^{n}\sum^{n}\frac{y_if_if_jy_l}{v_i^3v_j^2v_l}+$$

$$\left.+\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{f_if_jy_j^2}{v_i^2v_j^4}+\sum_{\substack{i=1j=1l=1\\ i\neq j,l}}^{n}\sum^{n}\sum^{n}\frac{f_iy_j^2f_l}{v_i^2v_j^2v_l^2}+\sum_{\substack{i=1j=1l=1k=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum^{n}\sum^{n}\sum^{n}\frac{f_iy_jf_ly_k}{v_i^2\ v_j\ v_l^2\ v_k}\right]$$

3), 4), 5), 6), 7) e 8) $\longrightarrow 0$

9)

$$5a^2\sigma^6\left[3a^2\sum_{i=1}^{n}\frac{y_i^2}{v_i^4}-6a\sum_{i=1}^{n}\frac{y_i^2}{v_i^5}-6a\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_iy_j}{v_i^4v_j}+3a^2\sum_{i=1}^{n}\frac{y_i^2}{v_i^6}+\right.$$

$$\left.+6\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_iy_j}{v_i^5v_j}+3\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_j^2}{v_i^4v_j^2}+3\sum_{\substack{i=1j=1l=1\\ i\neq j,l}}^{n}\sum^{n}\sum^{n}\frac{y_jy_l}{v_i^4v_jv_l}\right]$$

10)

$$6a^2\sigma^6\left[a^2\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i^2}{v_i^3v_j}-2a\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i^2}{v_i^4v_j}-2a\sum_{\substack{i=1j=1l=1\\ i\neq j,l}}^{n}\sum^{n}\sum^{n}\frac{y_iy_j}{v_i^3v_jv_l}+\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_i^2}{v_i^5v_j}+\right.$$

$$\left.+2\sum_{\substack{i=1j=1l=1\\ i\neq j,l}}^{n}\sum^{n}\sum^{n}\frac{y_iy_j}{v_i^4v_jv_l}+\sum_{\substack{i=1j=1l=1\\ i\neq j,l}}^{n}\sum^{n}\sum^{n}\frac{y_j^2}{v_i^3v_j^2v_l}\sum_{\substack{i=1j=1l=1k=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\sum^{n}\sum^{n}\sum^{n}\frac{y_jy_l}{v_i^3\ v_j\ v_l\ v_k}\right]$$

11)

$$30a\sigma^6\left[a^2\sum_{i=1}^{n}\frac{y_i^2}{v_i^5}-a\sum_{i=1}^{n}\frac{y_i^2}{v_i^6}+2a\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{y_iy_j}{v_i^5v_j}-\sum_{i=1}^{n}\frac{y_i^2}{v_i^7}-\right.$$

$$-2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i y_j}{v_i^6 v_j} - \sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_j^2}{v_i^5 v_j^2} - \sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_j y_l}{v_i^5 v_j v_l}\Bigg]$$

12)

$$12a\sigma^6\left[-a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i^2}{v_i^4 v_j} + 2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i^2}{v_i^5 v_j} + 2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_i y_j}{v_i^4 v_j v_l} - \sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_j^2}{v_i^4 v_j^2 v_l}\right] -$$

$$-12a\sigma^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i^2}{v_i^6 v_j} - 12a\sigma^4\left[2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_i y_j}{v_i^5 v_j v_l} + \sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\frac{y_j y_l}{v_i^4\ v_j\ v_l\ v_k}\right]$$

13)

$$12a\sigma^6\left[-2a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i y_j}{v_i^3 v_j^2} + 2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i y_j}{v_i^4 v_j^2} - 2a\,(1-2a)\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_j^2}{v_i^3 v_j^3} +\right.$$

$$+3a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_i y_j}{v_i^3 v_j^2 v_l} - 4a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_i y_j}{v_i^4 v_j^3} - 8\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_i y_j}{v_i^4 v_j^2 v_l} -$$

$$\left.-2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{y_j^2}{v_i^3 v_j^4} - 2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_j^2}{v_i^3 v_j^2 v_l^2} - 24\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\frac{y_j y_k}{v_i^3\ v_j\ v_l^2\ v_k}\right]$$

14)

$$12a\sigma^4\left[2a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_j y_l}{v_i v_j^3 v_l} - (1-2a)\,\sigma^4\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_l^2}{v_i v_j^2 v_l^3}\right] +$$

$$12a\sigma^6\left[-5a^2\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_j y_l}{v_i v_j^2 v_l} + 10a\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\frac{y_j y_k}{v_i\ v_j^2\ v_l^2\ v_k} + 10\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_j y_l}{v_i v_j^3 v_l^3} +\right.$$

$$-20\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\frac{y_j y_k}{v_i\ v_j^3\ v_l^2\ v_k} - 5\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j,l}}^{n}\frac{y_l^2}{v_i v_j^2 v_l^4} - 5\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j,l,k\ ;\ j\neq l}}^{n}\frac{y_l^2}{v_i\ v_j^2\ v_l^2\ v_k^2} -$$

$$-5\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{t=1}^{n}\frac{y_l y_t}{v_i\ v_j^2\ v_l\ v_k^2\ v_t}\Bigg]$$
$$i\neq j,l,k,t\ ;\ j\neq l,k\ ;\ l\neq k$$

15)

$$45\sigma^6\left[a^2\sum_{i=1}^{n}\frac{y_i^2}{v_i^6}-2a\sum_{i=1}^{n}\frac{y_i^2}{v_i^7}-2a\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i y_j}{v_i^6 v_j}+\sum_{i=1}^{n}\frac{y_i^2}{v_i^8}+2\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i y_j}{v_i^7 v_j}+\right.$$
$$i\neq j \qquad\qquad i\neq j$$

$$\left.+\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_j^2}{v_i^6 v_j^2}+\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j y_l}{v_i^6 v_j v_l}\right]$$
$$i\neq j \qquad\qquad i\neq j,l$$

16)

$$10\sigma^6\left[3a^2\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i^2}{v_i^4 v_j^2}-4a\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i^2}{v_i^5 v_j^2}-6a\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j}{v_i^4 v_j v_l^2}+6\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j}{v_i^5 v_j v_l^2}+\right.$$
$$i\neq j \qquad\qquad i\neq j,l \qquad\qquad i\neq j,l$$

$$\left.+3\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j^2}{v_i^4 v_j^2 v_l^2}\right]+6\sigma^4\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i^2}{v_i^6 v_j^2}+2\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_l y_k}{v_i^4\ v_j^2\ v_l\ v_k}$$
$$i\neq j \qquad\qquad i\neq j,l,k\ ;\ j\neq l$$

17)

$$120\sigma^6\left[a^2\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i y_j}{v_i^4 v_j^2}-2a\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i y_j}{v_i^5 v_j^2}+(1-2a)\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_j^2}{v_i^4 v_j^3}-\right.$$
$$i\neq j \qquad\qquad i\neq j \qquad\qquad i\neq j$$

$$-2a\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j}{v_i^4 v_j^2 v_l}+2\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_i y_j}{v_i^5 v_j^3}+4\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i y_j}{v_i^5 v_j^2 v_l}+$$
$$i\neq j,l \qquad\qquad i\neq j \qquad\qquad i\neq j,l$$

$$\left.+\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\frac{y_j^2}{v_i^4 v_j^4}+\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{k=1}^{n}\frac{y_j y_k}{v_i^4\ v_j\ v_l^2\ v_k}\right]+24\sigma^4\sum_{\substack{i=1}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j^2}{v_i^4 v_j^2 v_l^2}$$
$$i\neq j \qquad\qquad i\neq j,l,k\ ;\ j\neq l \qquad\qquad i\neq j,l$$

18) = 14) substituindo $\frac{-4a}{v_i}$ por $\frac{6}{v_i^2}$

19) = 16) substituindo $\frac{1}{v_j^2}$ por $\frac{2}{v_j}$

20) = 0

O termo IV de Δ^4 é

$$-4\theta\chi^3 = -\frac{4}{d^3}\tilde{\beta}_1^3\theta\left(\tilde{d}-d\right)^3 =$$

$$= \left(-\frac{4}{d^3}\tilde{\beta}_1^3\right)\frac{1}{d}\left[a\sum_{i=1}^{n}\frac{\varepsilon_i y_i}{v_i} - \sum_{i=1}^{n}\frac{\varepsilon_i y_i}{v_i^2} - \sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{\varepsilon_i y_j}{v_i v_j}\right]\left(\tilde{d}-d\right)^3 =$$

$$= \left(-\frac{4}{d^3}\tilde{\beta}_1^3\right)\left[\underbrace{\frac{1}{d}a\sum_{i=1}^{n}\frac{\varepsilon_i y_i}{v_i}\left(\tilde{d}-d\right)^3 - \frac{1}{d}\sum_{i=1}^{n}\frac{\varepsilon_i y_i}{v_i^2}\left(\tilde{d}-d\right)^3}_{IV/1 \qquad IV/2} - \underbrace{\frac{1}{d}\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{\varepsilon_i y_j}{v_i v_j}\left(\tilde{d}-d\right)^3}_{IV/3}\right]$$

Necessitamos apenas de calcular o valor médio destes três sub-termos de IV.

$$\mathrm{E}\left(IV/1\right) = \frac{1}{d}\left[24a^4\sigma^4\sum_{i=1}^{n}\frac{f_i}{v_i^4} + 24a^4\sigma^4\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i f_j^2}{v_i^2 v_j^2} + 30a^4\sigma^6\sum_{i=1}^{n}\frac{y_i f_i}{v_i^4} +\right.$$

$$+6a^4\sigma^6\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^2 v_j^2} + 12a^4\sigma^6\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^3 v_j} + 2a^4\sigma^6\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i f_i}{v_i^2 v_j v_l} - 60a^3\sigma^6\sum_{i=1}^{n}\frac{y_i f_i}{v_i^5} -$$

$$-12a^3\sigma^6\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^2 v_j^3} - 12a^3\sigma^6\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^4 v_j} - 12a^3\sigma^6\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_j f_j}{v_i^2 v_j^3} -$$

$$-4a^3\sigma^6\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i f_i}{v_i^2 v_j^2 v_l} - 24a^3\sigma^6\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i y_j}{v_i^3 v_j^2} - 24a^3\sigma^6\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_j}{v_i^3 v_j^2} -$$

$$-4a^3\sigma^6\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_j y_l}{v_i v_j^2 v_l^2} + 30a^2\sigma^6\sum_{i=1}^{n}\frac{y_i f_i}{v_i^6} + 6a^2\sigma^6\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^2 v_j^4} + 6a^2\sigma^6\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^4 v_j^2} +$$

$$\left.+6a^2\sigma^6\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_j f_j}{v_i^2 v_j^4} + 2a^3\sigma^6\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i f_i}{v_i^2 v_j^2 v_l^2} + 24a^2\sigma^6\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i y_j}{v_i^4 v_j^2} +\right.$$

$$+24a^2\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_j}{v_i^4 v_j^2}+12a^2\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_j y_l}{v_i^2 v_j^2 v_l^2}+12a^2\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j f_l}{v_i^2 v_j^2 v_l^2}+$$

$$+12a^2\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^4 v_j}+12a^2\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_j f_j}{v_i^2 v_j^3}+4a^2\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i y_i}{v_i^2 v_j^2 v_l}+60a^4\sigma^6\sum_{i=1}^{n}\frac{y_i f_i}{v_i^4}+$$

$$+24a^4\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_j f_j}{v_i v_j^3}+12a^4\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^2 v_j^2}+4a^4\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j f_j}{v_i v_j^2 v_l}-60a^3\sigma^6\sum_{i=1}^{n}\frac{y_i f_i}{v_i^5}+$$

$$-12a^3\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_j f_j}{v_i v_j^4}-12a^3\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^3 v_j^2}-12a^3\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^2 v_j^3}-$$

$$-4a^3\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j f_j}{v_i v_j^2 v_l^2}-24a^3\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i y_j}{v_i^3 v_j^2}-24a^3\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i y_j}{v_i^2 v_j^3}-$$

$$-8a^3\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_j y_l}{v_i v_j^2 v_l^2}+60a^3\sigma^6\sum_{i=1}^{n}\frac{y_i f_i}{v_i^5}+12a^3\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_j f_j}{v_i^2 v_j^3}+12a^3\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^4 v_j}+$$

$$+12a^3\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^2 v_j^3}+4a^3\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j f_j}{v_i^2 v_j^2 v_l}-60a^2\sigma^6\sum_{i=1}^{n}\frac{y_i f_i}{v_i^6}-24a^3\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_j f_j}{v_i^2 v_j^4}-$$

$$-12a^2\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^2 v_j^4}-4a^2\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_j f_j}{v_i^2 v_j^2 v_l^2}-24a^3\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i y_j}{v_i^4 v_j^2}-$$

$$-24a^2\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i y_j}{v_i^2 v_j^4}-8a^2\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_j y_l}{v_i^2 v_j^2 v_l^2}+24a^3\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i y_j}{v_i^3 v_j^2}+$$

$$+24a^3\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i y_j}{v_i^2 v_j^3}+8a^3\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i y_l}{v_i^2 v_j v_l^2}-24a^2\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i y_j}{v_i^4 v_j^2}-$$

$$-8a^3\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i y_l}{v_i^2 v_j^2 v_l^2}-24a^2\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{f_i y_j}{v_i^2 v_j^4}-16a^2\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i y_l}{v_i^2 v_j^2 v_l^2}-$$

$$-48a^2\sigma^6\sum_{\substack{i=1\\i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i f_i}{v_i^4 v_j^2}-8a^2\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{f_i y_l}{v_i^2 v_j^2 v_l^2}-8a^2\sigma^6\sum_{\substack{i=1\\i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i f_i}{v_i^2 v_j^2 v_l^2}-$$

$$-12a^3\sigma^6\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i}{v_i^3v_j}+24a\sigma^6\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_l}{v_i^2v_j^2v_l^2}\Bigg]$$

$\mathrm{E}\,(IV/2) = \mathrm{E}\,(IV/1)$ substituindo $\frac{a}{v_i}$ por $-\frac{1}{v_i^2}$

$$\mathrm{E}\,(IV/3) = \sigma^4\left[24a^2\sum_{i=1}^{n}\frac{f_i^3}{v_i^4}+24a^4\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_if_if_j^2}{v_i^2v_j^2}\right]+$$

$$+\sigma^6\Bigg[30a^2\sum_{i=1}^{n}y_if_i\left(-\tfrac{1}{v_i^6}-\tfrac{2a}{v_i^5}+\tfrac{2}{v_i^4}+\tfrac{a^2}{v_i^4}\right)+18a^4\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}y_if_i\left(\tfrac{1}{v_i^2v_j^2}+\tfrac{2}{v_i^3v_j}\right)-$$

$$-12a^3\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}y_if_i\left(\tfrac{1}{v_i^4v_j}+\tfrac{1}{v_i^3v_j^2}+\tfrac{1}{v_i^2v_j^3}+\tfrac{2}{v_i^4v_j^2}\right)-$$

$$-6a^2\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}y_if_i\left(\tfrac{6}{v_i^4v_j^2}+\tfrac{2}{v_i^2v_j^4}-\tfrac{2}{v_i^4v_j}-\tfrac{2}{v_i^3v_j^2}\right)-24a^3\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}y_if_j\left(\tfrac{1}{v_i^2v_j^3}+\tfrac{1}{v_i^3v_j^2}+\tfrac{1}{v_i^2v_j^4}\right)-$$

$$-24a^2\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}y_if_j\left(\tfrac{1}{v_i^4v_j^2}\right)-12a^3\sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n}\frac{y_i}{v_i^3v_j}+6a^4\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_if_i}{v_i^4v_jv_l}+$$

$$+2a^3\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}y_if_i\left(\tfrac{1}{v_i^2v_j^2v_l^2}-\tfrac{2}{v_i^2v_j^2v_l}\right)-4a^2\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}y_if_i\left(\tfrac{3}{v_i^2v_j^2v_l^2}-\tfrac{1}{v_i^2v_j^2v_l}\right)-$$

$$-4a^3\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}y_if_j\left(\tfrac{2}{v_i^2v_j^2v_l^2}-\tfrac{1}{v_i^2v_j^2v_l}\right)-8a^3\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_if_j}{v_i^2v_j^2v_l^2}+$$

$$+24a\sum_{\substack{i=1\\ i\neq j\neq l}}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{y_i}{v_i^2v_j^2v_l^2}\Bigg]$$

Finalmente chegamos ao último termo de Δ^4, o termo V que é

$$\chi^4=\frac{\tilde{\beta}_1^4}{d^4}\left(\tilde{d}-d\right)^4=\frac{\tilde{\beta}_1^4}{d^4}\left(\tilde{d}-d\right)\left(\tilde{d}-d\right)^3=$$

$$= \frac{\tilde{\beta}_1^4}{d^4}\left[2a\sum_{i=1}^{n}\frac{f_i y_i}{v_i} + a\sum_{i=1}^{n}\frac{\varepsilon_i^2}{v_i} - \sum_{i=1}^{n}\frac{\varepsilon_i^2}{v_i^2} - \sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{\varepsilon_i\varepsilon_j}{v_i v_j}\right]\left(\tilde{d}-d\right)^3 =$$

$$= \frac{\tilde{\beta}_1^4}{d^4}\Bigg[\underbrace{2a\sum_{i=1}^{n}\frac{f_i y_i}{v_i}\left(\tilde{d}-d\right)^3}_{V/1} + \underbrace{a\sum_{i=1}^{n}\frac{\varepsilon_i^2}{v_i}\left(\tilde{d}-d\right)^3}_{V/2} - \underbrace{\sum_{i=1}^{n}\frac{\varepsilon_i^2}{v_i^2}\left(\tilde{d}-d\right)^3}_{V/3} -$$

$$\underbrace{-\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{\varepsilon_i\varepsilon_j}{v_i v_j}\left(\tilde{d}-d\right)^3}_{V/4}\Bigg]$$

Mais uma vez necessitamos de calcular o valor esperado destes quatro sub-termos.

$\mathrm{E}\,(V/1) = \mathrm{E}\,(IV/1)$ se substituirmos ay_i por $2af_i$.

$$\mathrm{E}\,(V/2) = \sigma^6\left\{a^3\left[180\sum_{i=1}^{n}\frac{f_i^2}{v_i^4} + 36\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n} f_i^2\left(\frac{1}{v_i^3 v_j} + \frac{1}{v_i^2 v_j^2}\right)\right] -\right.$$

$$\left. -a^2\left[180\sum_{i=1}^{n}\frac{f_i^2}{v_i^5} + 36\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i^2}{v_i^3 v_j^2} + 36\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i^2}{v_i^3 v_j^3} + 144\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{f_i f_j}{v_i^3 v_j^2}\right]\right\} +$$

$$+\sigma^8\left\{a^3\left[105\sum_{i=1}^{n}\frac{1}{v_i^4} + 27\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{1}{v_i^2 v_j^2} + 45\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{1}{v_i^3 v_j} + 9\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{1}{v_i^3 v_j v_l}\right] -\right.$$

$$-a^2\left[315\sum_{i=1}^{n}\frac{1}{v_i^5} + 126\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{1}{v_i^3 v_j^2} + 90\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{1}{v_i^4 v_j} + 18\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{1}{v_i^2 v_j^2 v_l} +\right.$$

$$\left.\left. +9\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{1}{v_i^3 v_j v_l}\right] + a\left[315\sum_{i=1}^{n}\frac{1}{v_i^6} + 45\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{1}{v_i^5 v_j} + 207\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{1}{v_i^4 v_j^2} +\right.\right.$$

$$+162\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^3v_j^3}+27\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{1}{v_i^2v_j^2v_l^2}+12\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{1}{v_i^2v_j^2v_l}+42\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{1}{v_i^3v_j^2v_l}\Bigg]-$$

$$-\Bigg[105\sum_{i=1}^{n}\frac{1}{v_i^7}+30\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^5v_j}+165\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^5v_j^2}+99\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^4v_j^3}+48\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^4v_j^2}+$$

$$+18\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^3v_j^3}+117\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{1}{v_i^3v_j^2v_l^2}+12\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{1}{v_i^3v_j^2v_l}+6\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{1}{v_i^2v_j^2v_l^2}\Bigg]\Bigg\}$$

$\mathrm{E}\,(V/3) = \mathrm{E}\,(V/2)$ substituindo $\frac{a}{v_i}$ por $-\frac{1}{v_i^2}$

$$\mathrm{E}\,(V/4)=\sigma^6\Bigg\{a^3\Bigg[144\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{f_if_j}{v_i^3v_j^2}+24\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{f_if_j}{v_i^2v_j^2v_l}\Bigg]-a^2\Bigg[144\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{f_if_j}{v_i^4v_j^2}+$$

$$+48\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{f_i^2}{v_i^4v_j^2}+120\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{f_if_j}{v_i^2v_j^2v_l^2}+24\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{f_i^2}{v_i^2v_j^2v_l^2}\Bigg]\Bigg\}+$$

$$+\sigma^8\Bigg\{-a^2\Bigg[180\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^4v_j^2}+108\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^3v_j^3}+24\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{1}{v_i^2v_j^2v_l^2}+$$

$$+48\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{1}{v_i^3v_j^2v_l}+18\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{1}{v_i^2v_j^2v_l^2}+6\sum_{\substack{i=1j=1l=1k=1\\ i\neq j\neq l\neq k}}^{n}\sum^{n}\sum^{n}\sum^{n}\frac{1}{v_i^2\ v_j^2\ v_l\ v_k}\Bigg]+$$

$$+a\Bigg[120\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^6v_j^2}+144\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^4v_j^3}+240\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^5v_j^2}+24\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^4v_j^2}+$$

$$+72\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{1}{v_i^4v_j^2v_l}++300\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{1}{v_i^3v_j^2v_l^2}+24\sum_{\substack{i=1j=1l=1\\ i\neq j\neq l}}^{n}\sum^{n}\sum^{n}\frac{1}{v_i^3v_j^3v_l}+$$

$$+28\sum_{\substack{i=1j=1l=1k=1\\ i\neq j\neq l\neq k}}^{n}\sum^{n}\sum^{n}\sum^{n}\frac{1}{v_i^2\ v_j^2\ v_l^2\ v_k}\Bigg]-\Bigg[180\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^6v_j^2}+108\sum_{\substack{i=1j=1\\ i\neq j}}^{n}\sum^{n}\frac{1}{v_i^4v_j^4}+$$

$$+72\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{1}{v_i^4 v_j^3}+402\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{1}{v_i^4 v_j^2 v_l^2}+24\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{1}{v_i^4 v_j^2 v_l}+24\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1\\ i\neq j\neq l}}^{n}\frac{1}{v_i^3 v_j^2 v_l^2}+$$

$$+38\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j\neq l\neq k}}^{n}\frac{1}{v_i^2\ v_j^2\ v_l^2\ v_k^2}+4\sum_{\substack{i=1}}^{n}\sum_{\substack{j=1}}^{n}\sum_{\substack{l=1}}^{n}\sum_{\substack{k=1\\ i\neq j\neq l\neq k}}^{n}\frac{1}{v_i^2\ v_j^2\ v_l^2\ v_k}\Bigg]\Bigg\}$$

O valor esperado de Δ^4 resulta na soma de todos estes termos.

Apêndice C

Gráficos

Neste apêndice apresentamos os gráficos presentes no Capítulo 3, por forma a tornar mais fácil a sua consulta.

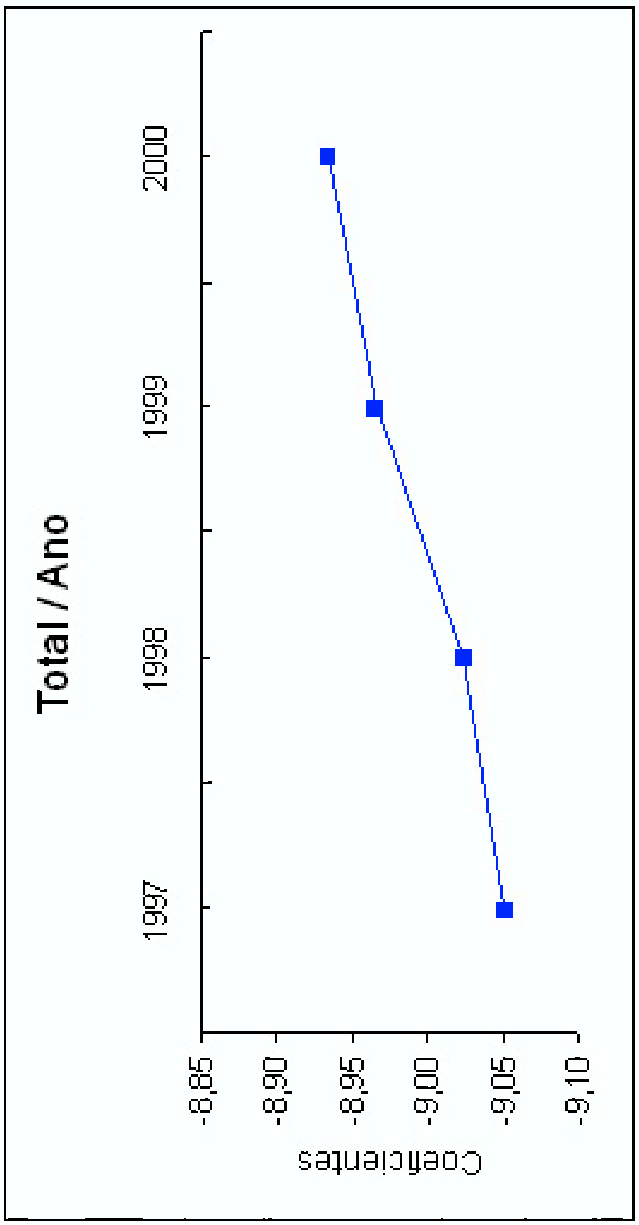

Figura 3.1 - Estimativas para o factor temporal - Total.

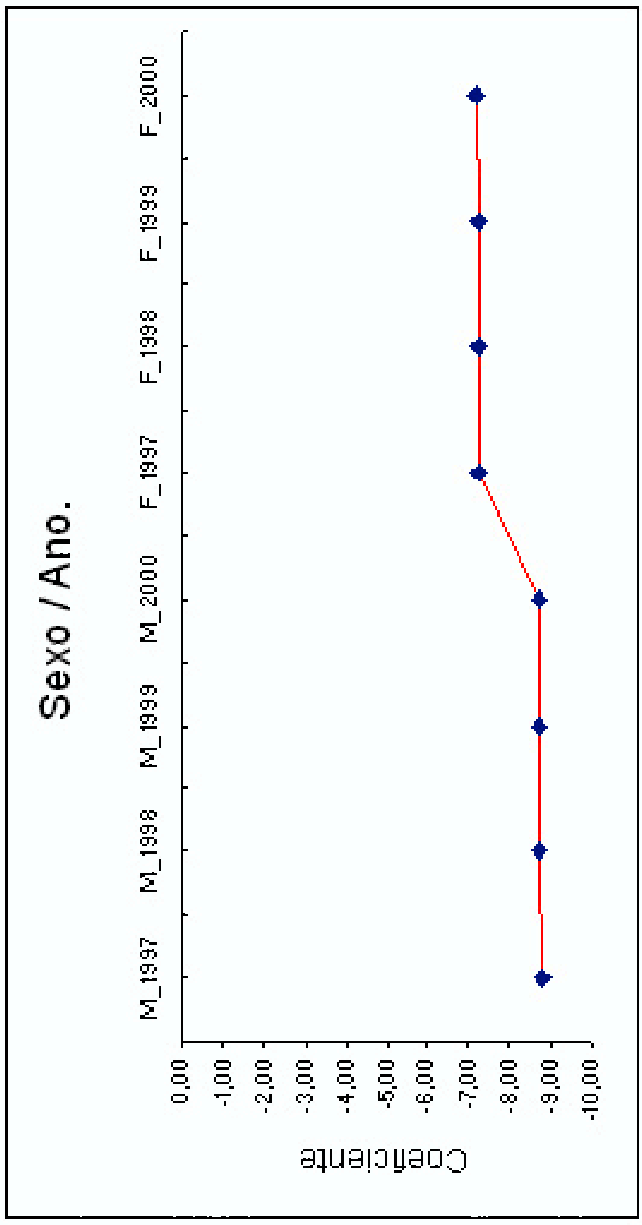

Figura 3.2 - Estimativas para o factor temporal - Sexo.

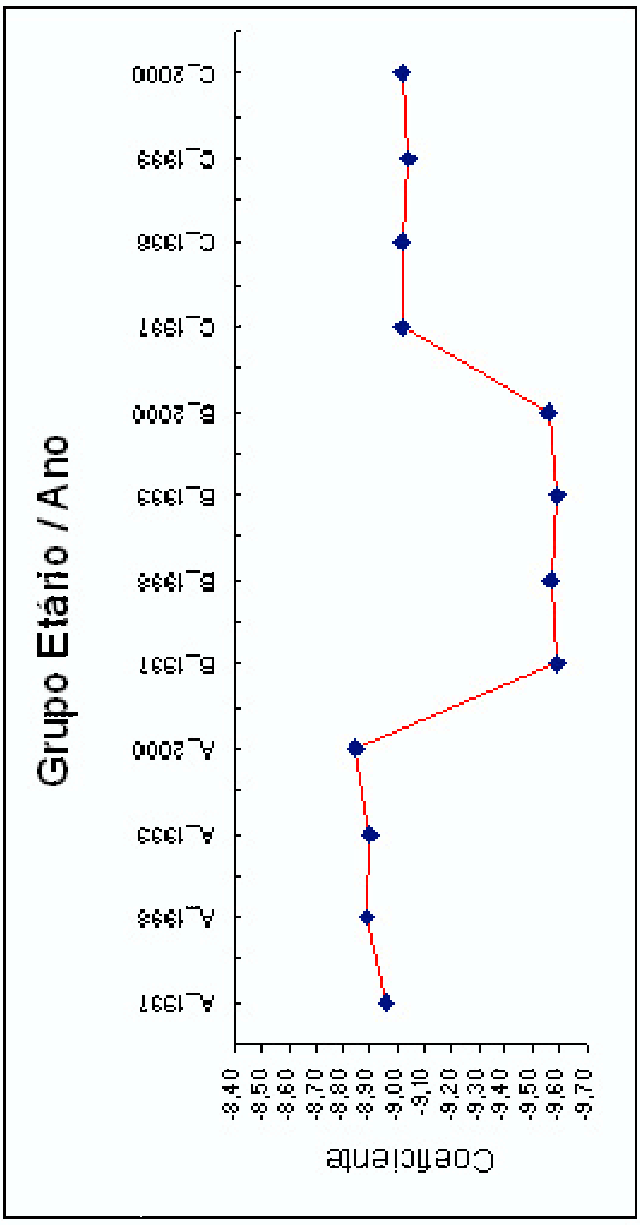

Figura 3.3 - Estimativas para o factor temporal - Grupo Etário.

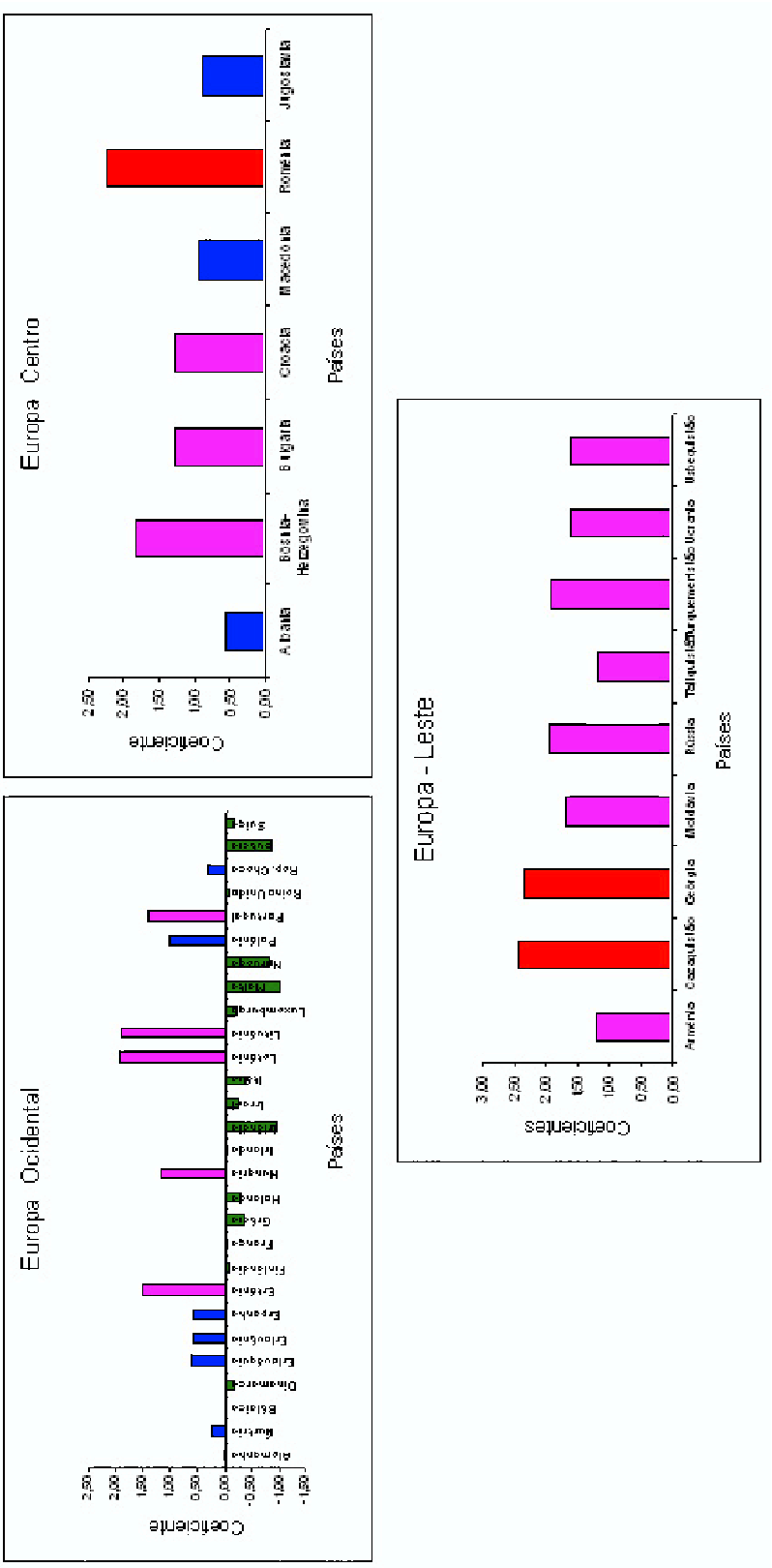

Figura 3.4 - Estimativas para o factor de localização - Total - WHO European Region.

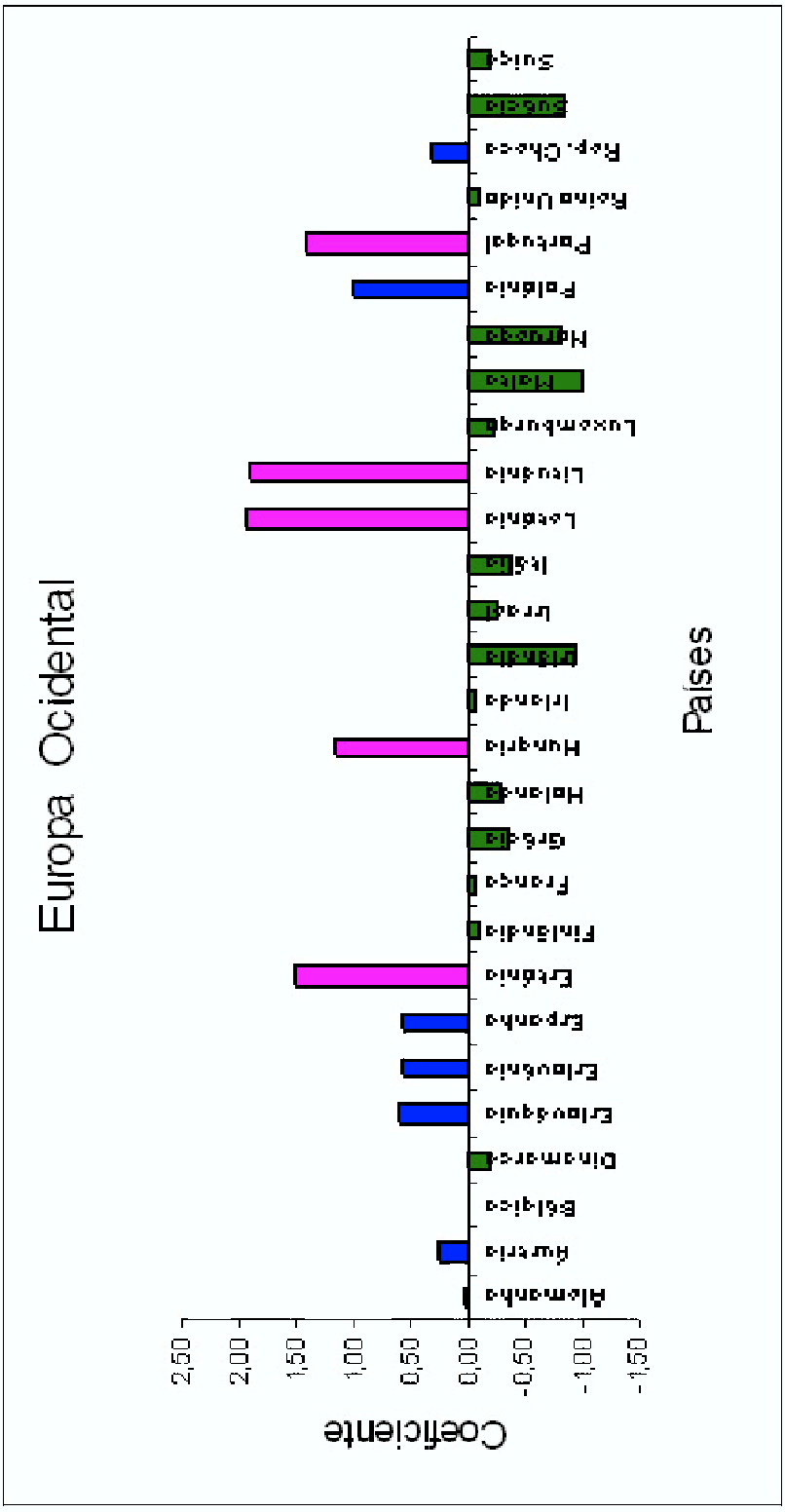

Figura 3.4 - Estimativas para o factor de localização - Total - Europa Ocidental.

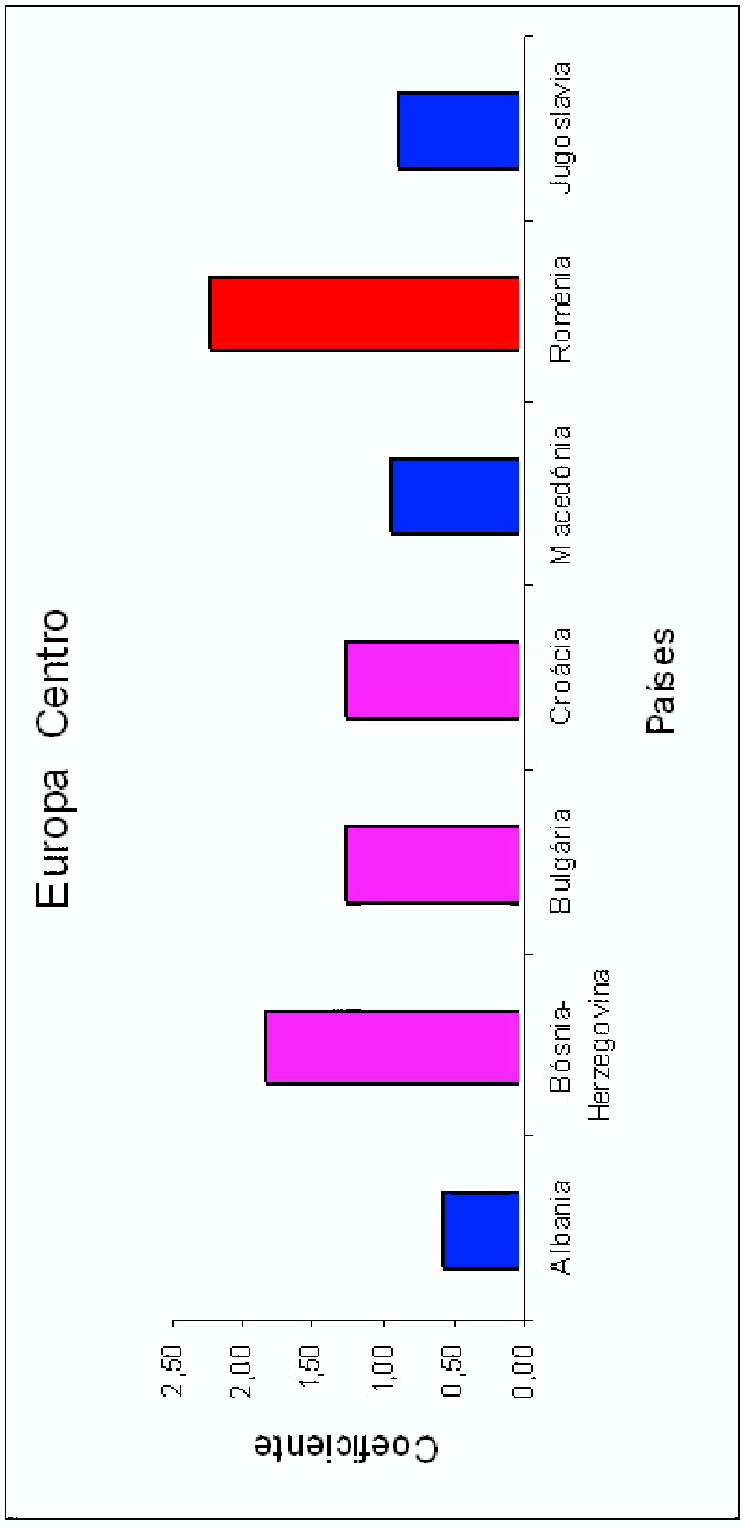

Figura 3.4 - Estimativas para o factor de localização - Total - Europa Central.

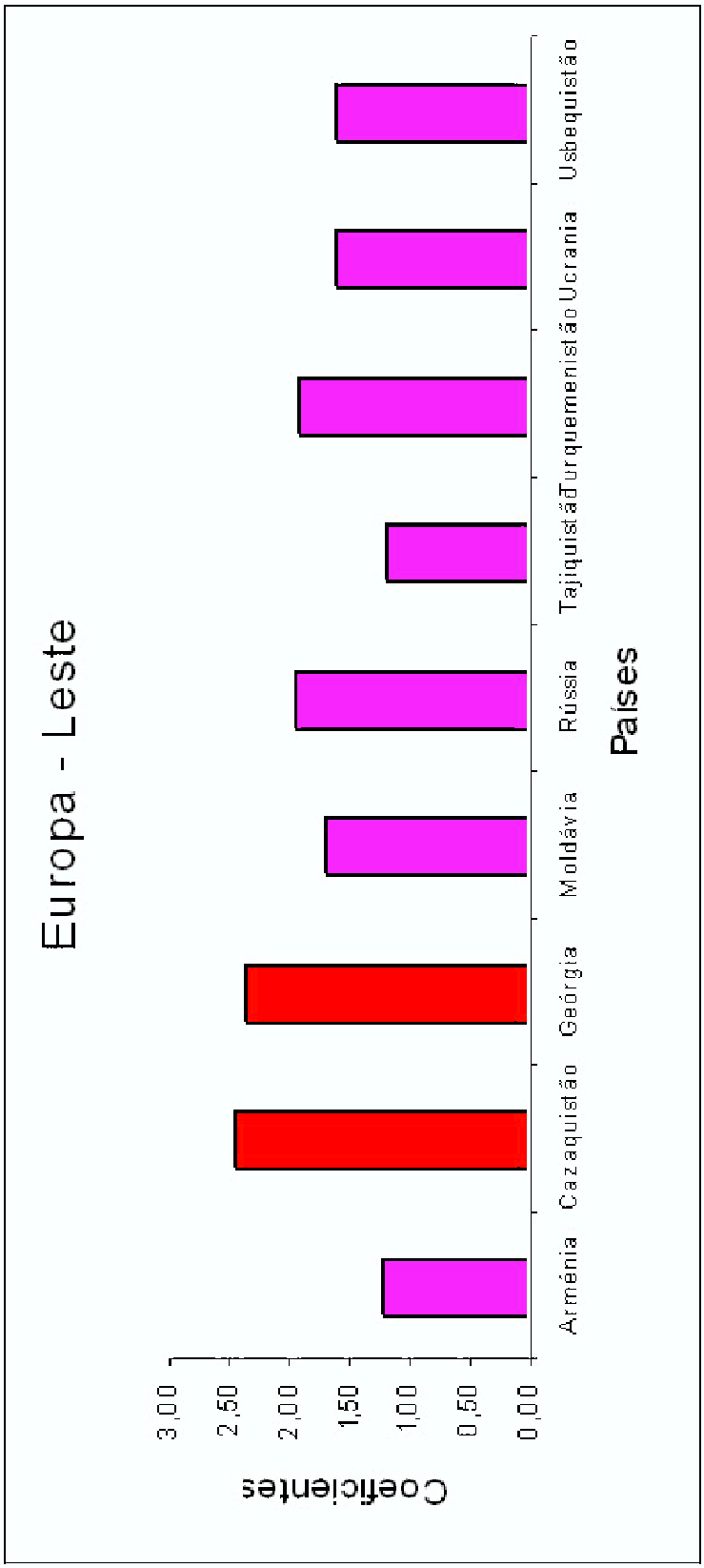

Figura 3.4 - Estimativas para o factor de localização - Total - Europa de Leste.

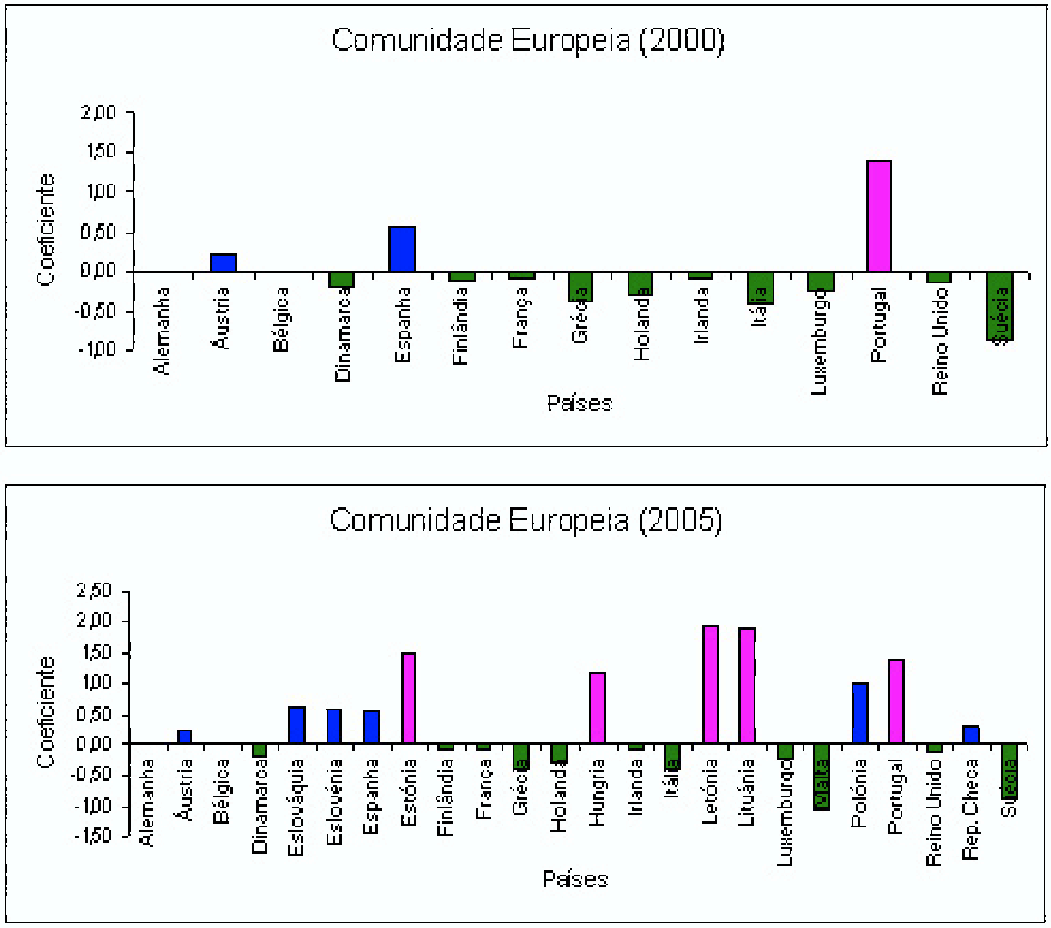

Figura 3.5 - Estimativas para o factor de localização - Total - Comunidade Europeia (2000 e 2005).

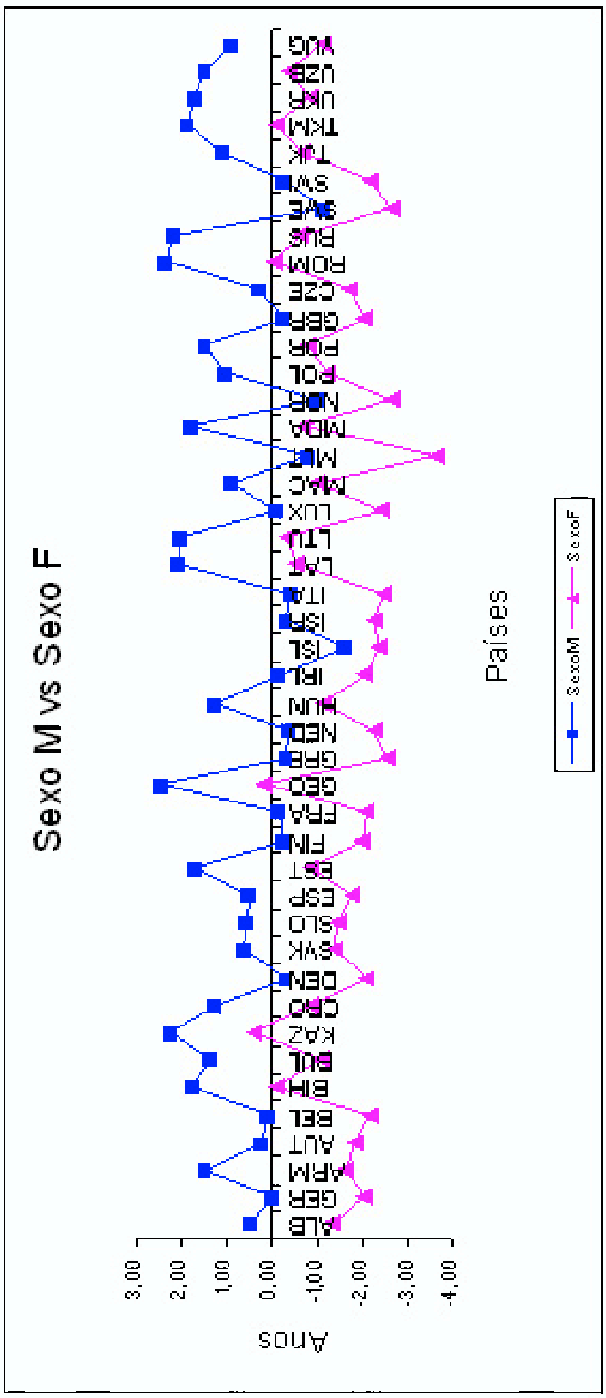

Figura 3.6 - Estimativas para o factor de localização - Sexo.

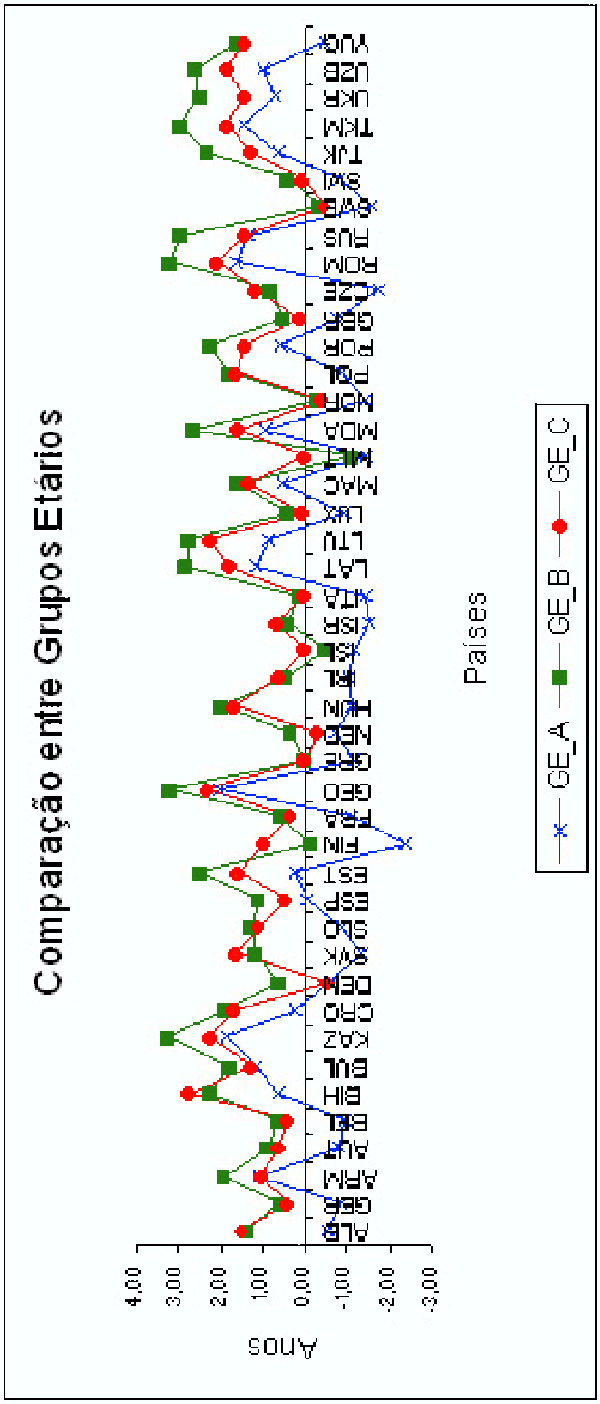

Figura 3.7 - Estimativas para o factor de localização - Grupos Etários.

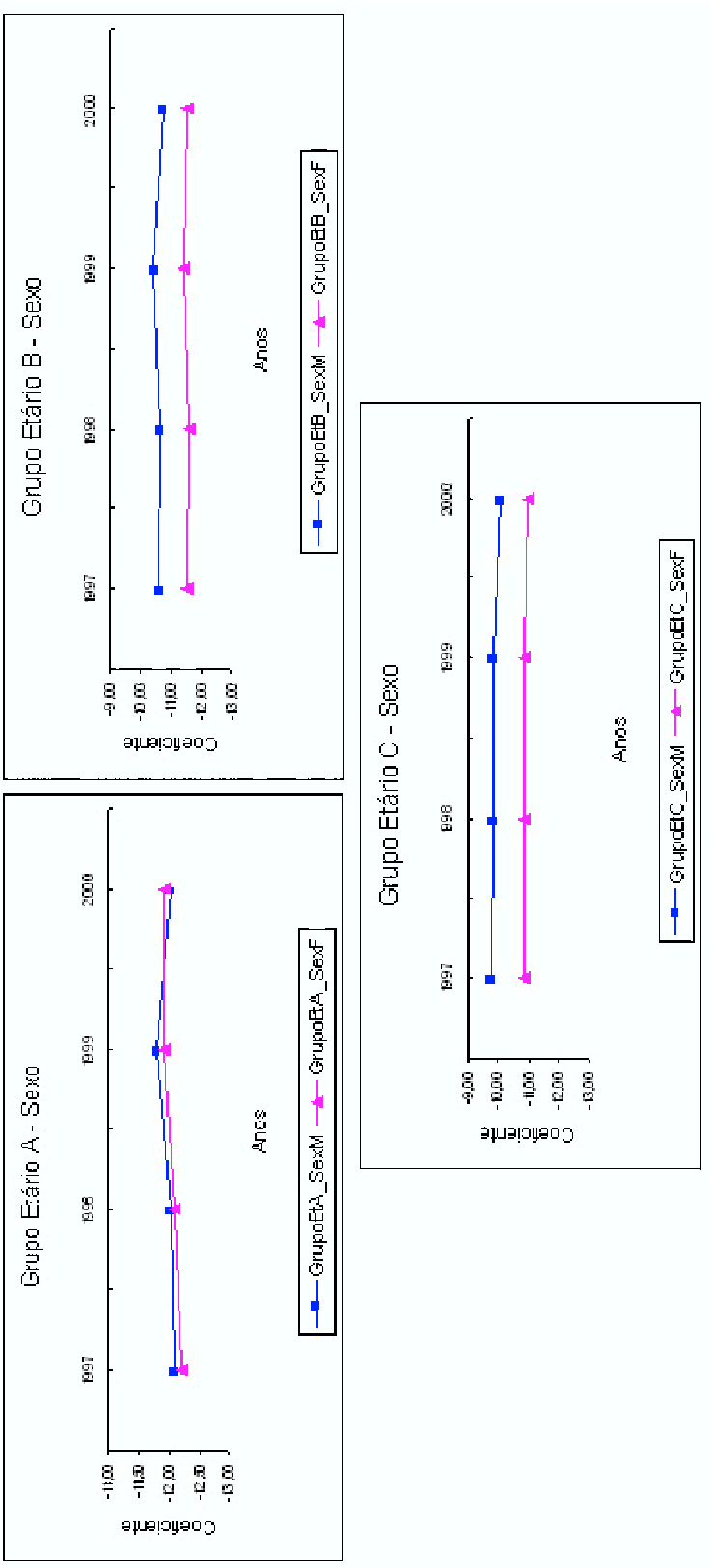

Figura 3.8 - Estimativas para o factor temporal - Grupos Etários / Sexo.

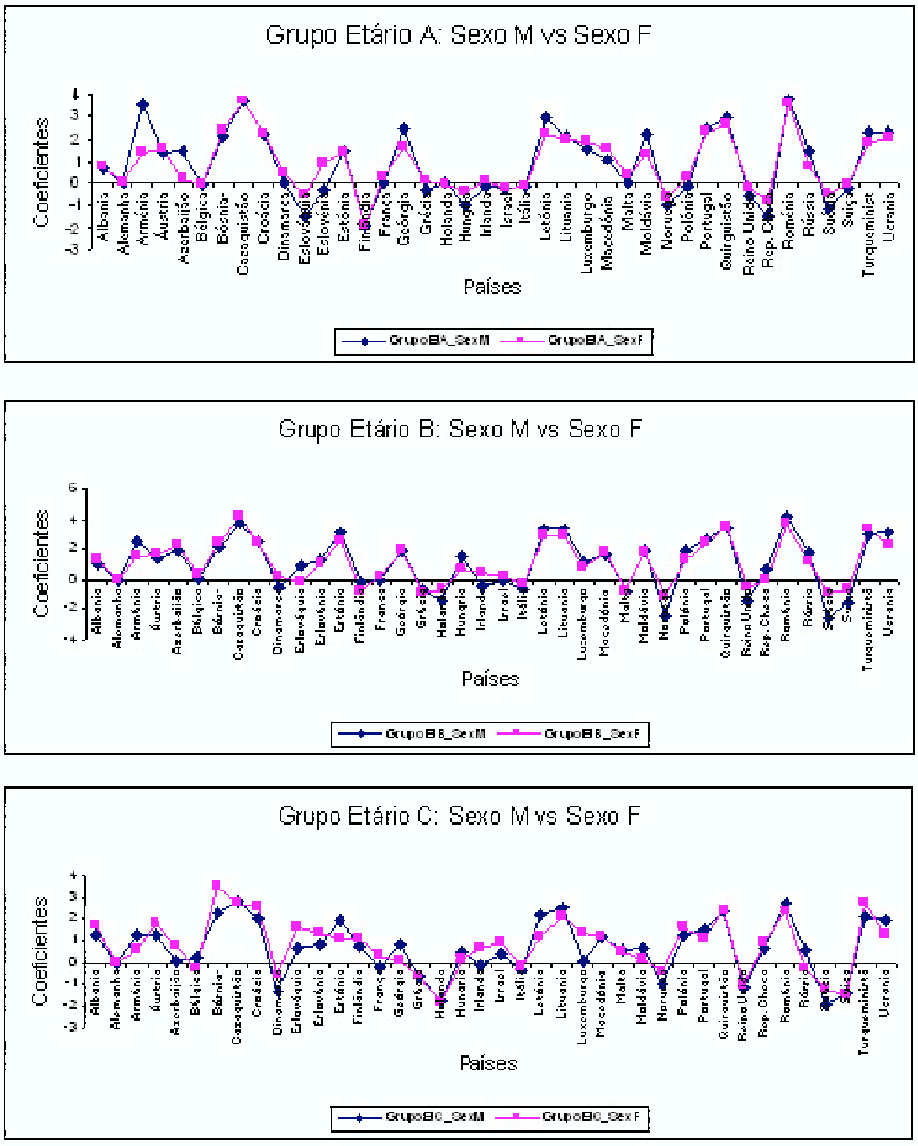

Figura 3.9 - Estimativas para o factor de localização - Grupos Etários / Sexo.

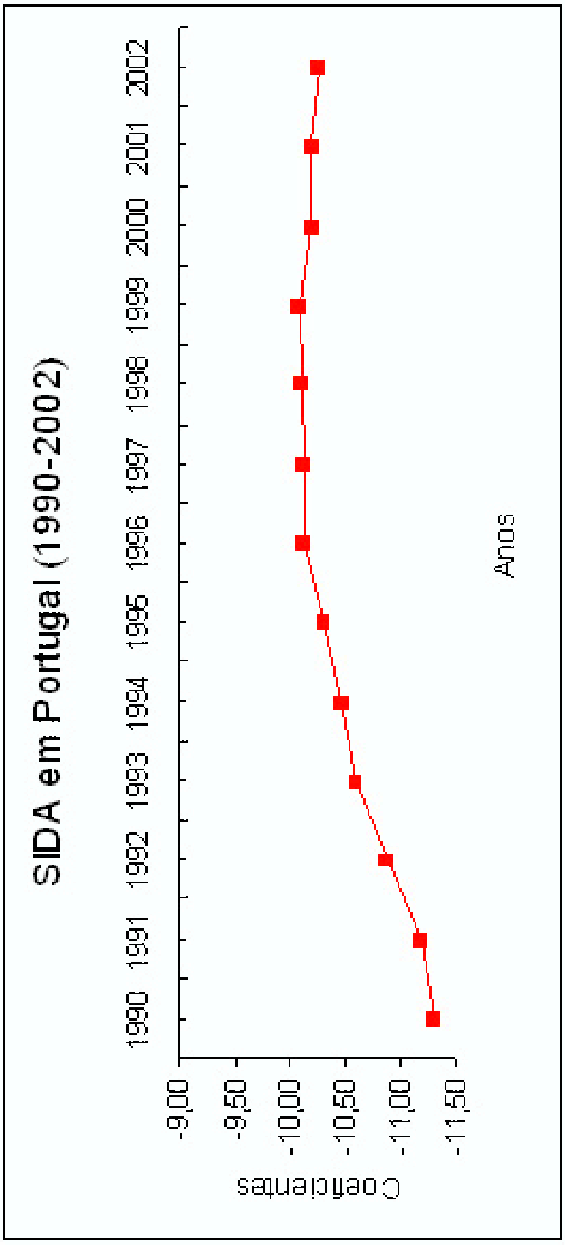

Figura 3.10 - Estimativas para o factor temporal - SIDA.

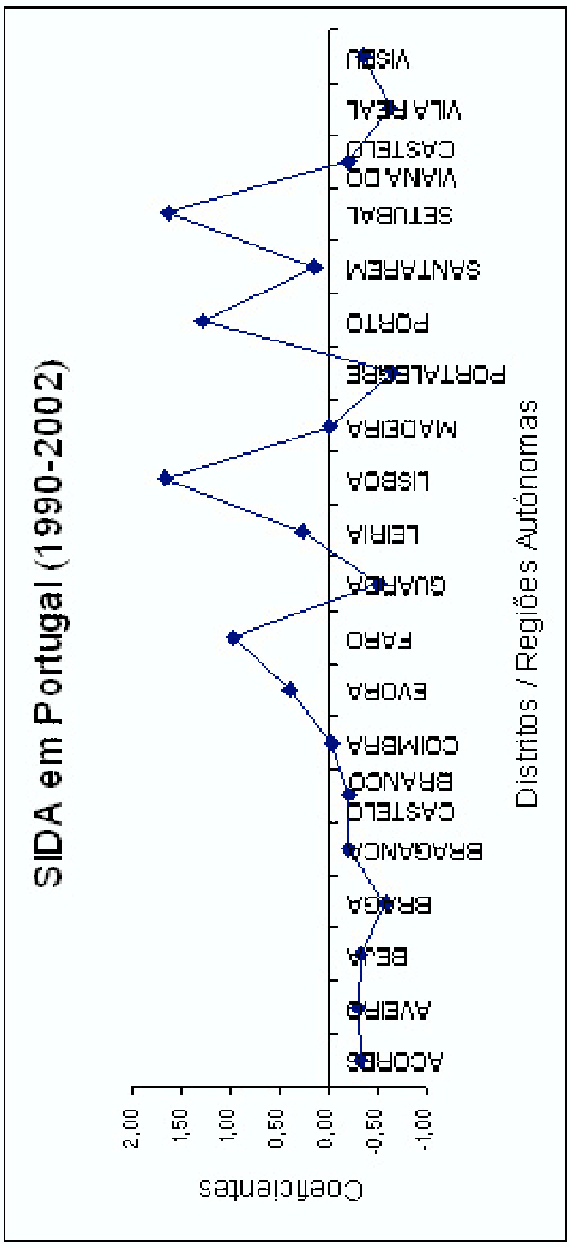

Figura 3.11 - Estimativas para o factor de localização - SIDA.

Bibliografia

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Aitchison, J. and Brown, J.A.C. (1957). *The LogNormal Distribution*. Number 5 in University Cambridge, Department of Applied Economics Monographs. Cambridge University Press.

[3] Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Sciences* **10**, p. 364-376.

[4] Amaral, J.A. (2000). Projecções sobre a incidência e prevalência de SIDA em Portugal. *Sociedade Portuguesa de Estatística. Boletim Informativo*, 1/2000, p. 21-24.

[5] Amemiya, T. (1981). Qualitative Response models: a survey. *Journal of Economic Literature* **19**, p. 1483-1536.

[6] Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press Cambridge, Massachusetts.

[7] Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions*. New York: Dover.

[8] Bateman, H. and Erdélyi, A. (1954). *Tables of Integral Transforms*. McGraw-Hill, new York, Toronto, London.

[9] Bapat, R.B. (2000). *Linear Algebra and Linear Models*. 2nd edition. Springer-Verlag, New York.

[10] Barndorff-Nielsen, O.E. and Cox, D.R. (1989). *Asymptotic Techniques for use in Statistics*. Chapman and Hall.

[11] Ben-Akiva, M., e Lerman, S.R. (1985). *Discrete Choice Analysis: Theory and Apllication to Travel demand*. The MIT Press, Cambridge, Mass.

[12] Berkson, J. (1944). Application of the Logistic Function to Bio-assay. *Journal of the American Statistical Association*, **9**,p. 357-365.

[13] Berkson, J. (1951). Why I Prefer Logits to Probits. *Biometrics*, **7**,p. 327-339.

[14] Berkson, J. (1980). Minimum Chi-Square, Not Maximum Likelihood!. *Annals of Mathematical Statistics*, **8**,p. 457-487.

[15] Bernardeau, F. and Kofman, L. (1995). Properties of Cosmological Density Distribution Function. *Astrophysical Journal*, Part **1**, Vol. **443**, No. **2**, p. 479-498.

[16] Blimmikov, S. and Moessner, R. (1998). Expansions for nearly Gaussian distributions. *Astronomy and Astrophysics Supplement Series* **130**, p. 193-205.

[17] Bliss, C.I. (1934-A). The Method of Probits. *Science* **79**, p. 38-39.

[18] Bliss, C.I. (1934-B). The Method of Probits. *Science* **79**, p. 409-410.

[19] Bonferroni, C.E. (1935). Il Calcolo delle Assicurazioni su Gruppi di Teste. In *Studii in Onore del Professore Salvatore Ortu Carboni*, p. 13-60. Rome.

[20] Bonferroni, C.E. (1936-). Teoria statistica delle classi e calcolo delle probabilità. Pubbllicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze **8**, p. 3-62.

[21] Breslow N.E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association* **91**, p. 14-28.

[22] Bretscher, O. (2004). *Linear Algebra with Applications*. 3rd edition. Prentice Hall.

[23] Câmara, V.M. e Tambellini, A.T. (2003). Considerações sobre o uso da epidemiologia nos estudos em saúde ambiental. *Revista Brasileira de Epidemilologia*, Vol.**6** No.**2**, p.95-104. São Paulo.

[24] Campbell, S.L. and Meyer,Jr., C.D. (1979). *Generalized Inverses of Linear Transformations*. New York: Dover Publications.

[25] Carvalho, L. e Diamantino, F. (1993). Epidemia de Sida em Portugal. Projecções pelo método de Back-Calculation. Em *A Estatística e o Futuro da Estatística*. (Pestana D., Turkman A., Branco J., Duarte L. e Pires A., eds.), p. 19-29. Edições SPE.

[26] Chen, C. S. and Savits, T.H. (1993). Some remarks on compound nonhomogeneous Poisson processes. *Statistics and Probability Letters* **17**, p. 179-187.

[27] Christensen, R. (1987). *Plane Answers to Complex Questions - The Theory of Linear Models*. Springer Verlag.

[28] Christensen, R. (1996). *Analysis of Variance, Design and Regression - Applied Statistical Methods*. Chapman and Hall/CRC.

[29] Constantine, G.M. (1987). *Combinatorial Theory and Statistical Design*. Wiley, New York.

[30] Constantine, G.M., and Savits, T.H. (1996). A Multivariate Faa di Bruno Formula with Applications. *Transactions of the American Mathematical Society* Vol. **348**, No. **2**, p. 503-520.

[31] Corte Real, P.A.R. (2001). *Modelos Lineares Normais com Conexão*. Dissertação apresentada para obtenção do Grau de Doutor em Matemática na Especialidade de Estatística pela Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia.

[32] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

[33] Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data* (Second ed.). Chapman and Hall, London. First edition, by Cox alone, in 1969.

[34] Crámer, H. (1957). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

[35] Cramer, J.S. (2002). The Origins of Logistic Regression. *Tinbergen Institute Discussion Paper*, TI 2002-199/4. Faculty of Economics and Econometrics, University of Amsterdam, and Tinbergen Institute.

[36] Cramer, J.S. (2003). *Logit Models from Economics and Other Fields*. Cambridge University Press.

[37] Davidson, R. and MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press.

[38] Dobson, A. (1990). *An Introduction to Generalized Linear Models*. Chapman and Hall, London.

[39] *Estatísticas Demográficas* (1990-2002). População e Condições Sociais. Estimativas da população residente em Portugal. Publicações do Instituto Nacional de Estatística.

[40] Faires J.D., Burden, R.L., Pirtle, B. and Sandberg, K. (2002). *Numerical Methods*. 3rd ed.. Brooks Cole.

[41] Fechner, G.T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel.

[42] Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, 3rd ed., Vol. **1**. Wiley, New York.

[43] Finney, D.(1971). *Probit Analysis* (3rd ed.). Cambridge: Cambridge University Press.

[44] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**, p. 399-433.

[45] Fisher, R.A. (1928). Moments and Product Moments of Sampling Distributions. *Procedures of London Mathematic Society* **30**, p. 199-238.

[46] Fisher, R.A. (1954). The Analysis of Variance with Various Binomial Transformations. *Biometrics* **10**, p. 130-151. With comments by M.S. Bartlett, F.J. Anscombe, W.G. Cochran and J. Berkson.

[47] Franses, P.H. and Paap, R. (2001). *Quantitative Models in Marketing Research.* Cambridge University Press, Cambridge.

[48] Frechet, M. (1940). Les Probabilitiés associees a un systeme d'événements compatibles et dependents. *Actualités scientifiques et industrielles*, No. **859**. Hermann and Cie, Paris.

[49] Gaddum, J.H. (1933). *Reports on Biological Standard III. Methods of Biological Assay Depending on a Quantal Response.* London: Medical Research Council. Special Report Series of the Medical Research Council, No. **183**.

[50] Galambos, J. (1975). Methods for proving Bonferroni Type Inequalities. *J. Lond. Math. Soc.*, **2**, p. 561-564.

[51] Galambos, J. (1977). Bonferroni Inequalities. *Annals of Probability*, **5**, p. 577-581.

[52] Gourieroux, C. (2000). *Econometrics of Qualitative Dependentent Variables.* Cambridge University Press, Cambridge. First edition (in French) in 1991.

[53] Hocking, R.R. (1996). *Methods and Applications of Linear Models - Regression and the Analysis of Variance.* John Wiley and Sons.

[54] Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression* (Second ed.). New York: Wiley. First edition in 1989.

[55] Howie, D. (2002). *Interpreting probability: controversies and developments in the early twentieth century.* Cambridge, UK: Cambridge University Press.

[56] Juszkiewicz, R., Weinberg, D.H., Amsterdamski, P., Chodorowski, R. and Bouchet, F. (1995). Weakly Nonlinear Gaussian Fluctuations and the Edgeworth Expansion. *Astrophysical Journal*, Part **1**, Vol. **442**, No. **1**, p. 39-56.

[57] Kelley, C.T. (2003). *Solving Nonlinear Equations with Newton´s Method (Fundamentals of Algorithms).* Society for Industrial and Applied Mathematic.

[58] Kendall, M.C. (1952). *The advanced theory of statistics, Vol.1, Distribution theory.* Fifth Edition. Charles Griffin and Co., London.

[59] Kendall Sc.D., M.G. (1961). *A Course in The Geometry of n Dimensions.* Being Number Eight of Griffin's Statistical Monographs and Courses. Charles Griffin and Company LTD, London.

[60] Kenney, J.F. and Keeping, E.S. (1951). *Mathematics of Statistics*, Pt.**2**, (2nd ed). Princeton, NJ: Van Nostrand.

[61] Kenney, J.F. and Keeping, E.S. (1962). The K-statistics. Sect. 7.9 in *Mathematics of Statistics*, Pt.**1**, (3rd ed.). Princeton, NJ: Van Nostrand.

[62] Liagre. J.B.J. (1852). *Calcul des probabilités et théorie des erreurs*. Bruxelles: Société pour l'emancipation intellectuelle (A.Jamard).

[63] Lukács, E. (1955). Applications of Faa di Bruno's Formula in Mathematical Statistics. *Am. Math. Monthly*, **62**, 340-348. MR 16:1037g.

[64] Lunneborg, C. (1994). *Modeling Experimental and Observational Data*. Duxbury Press.

[65] Luz, C., Matos, A. e Nunes, S. (2002). *Álgebra Linear*. Vol. 1. Escola Superior de Tecnologia de Setúbal.

[66] Lyapunov, A.M. (1954). *Collected Works 1*, Moscow, pp. 157-176.

[67] MacCullagh, P. and Nelder, J.H. (1989). *Generalized Linear Models* (Second ed.). London, UK: Chapman and Hall.

[68] Maddala, G.S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.

[69] Malthus, T.E. (1798). *An Essay on the principle of Population*. London: anon.

[70] McFadden, D.L., and Domencich, T.A. (1975, reprinted 1996). *Urban Travel Demand: A Behavioral Analysis*. North-Holland, Amesterdam, The Netherlands.

[71] McFadden, D.L. (1978). Modelling the Choice of Residential Location. In A. Karlquist et al.(ed.), *Spatial Interaction Theory and Residential Location*, p. 75-96. North-Holland, Amsterdam, The Netherlands.

[72] McFadden, D.L. (2001). Economic Choices. *American Economic Review* **91**, p. 352-370. Nobel prize acceptance speech.

[73] Menard, S. (1995). *Applied Logistic Regression Analysis*. Sage Publications, Thousands Oaks, California.

[74] Mexia, J.T. (1987). *Linear Restrictions on the Model and Scheffé's Multiple Comparison Method*. Trabalhos de Investigação, Nº3. Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa.

[75] Mexia, J.T. (1989). *Controlled Heteroscedasticity, Quotient Vector Spaces an F Tests for Hypothesis on Mean Vectors*. Trabalhos de Investigação, No.1. Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa.

[76] Mexia, J.T. (1990). *Variance Free Models*. Trabalhos de Investigação, No.2. Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa.

[77] Mexia, J.T. (1992). *Asymptotic Chi-Squared tests, Designs, and Log-Linear Models*. Trabalhos de Investigação, No.1. Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa.

[78] Mexia, J.T. (1995). *Introdução à Inferência Estatística Linear*. Centro de Estudos de Matemática Aplicada. Edições Universitárias Lusófonas.

[79] Mexia, J.T. e Nunes, S. (2003). The Collective Model Applied to Health Statistics. *Boletim do Instituto dos Actuários Portugueses*, No. **42**,p. 3-9.

[80] Mexia, J.T., Pereira, D. e Baeta, J. (1999). L_2 Environmental Indexes. *Listy Biometryczne - Biometrical Letters*, Vol. **38**, p. 33-40.

[81] Mexia, J.T., Pereira, D. e Baeta, J. (2001). Weighted Linear Joint Regression Analysis. *Listy Biometryczne - Biometrical Letters*, Vol. **36**, No. **2**, p. 137-143.

[82] Meyer, R.M. (1969). Note on a "Multivariate" Form of Bonferroni's inequalities. *Ann. Math. Statist.*, **40**, p. 692-693.

[83] Miner, J.R. (1933). Pierre-François Verhulst, the discoverer of the logistic curve. *Human Biology* **5**, p. 673-689.

[84] Moore, E.H. (1920). On the Reciprocal of General Algebraic Matrix. *Bulletin of the American Mathematical Society* **26**, p. 394-5.

[85] Moran, P.A.P. and Smith, C.A.B. (1966). *Commentary on R.A. Fisher's paper on the correlation between relatives on the supposition of Mendelian inheritance*. London: Galton Laboratory, University College London.

[86] *NIDI - The Netherlands Interdisciplinary Demographic Institute*: http://www.nidi.knaw.nl

[87] Nunes, S., Mexia,J.T. and Minder, C. (2003). Bias of Logits in Environmental Impact Studies. In: *Proceedings of the* 18^{th} *International Workshop on Statistical Modelling*, Verbeke, G., Molenberghs, G., Aerts, A. and Fieuws, S. (Eds.). Leuven: Katholieke Universiteit Leuven, p. 343-348.

[88] Nunes, S., Mexia,J.T. and Minder, C. (2004-A). Logit Model for Tuberculosis in Europe (1995-2000). In: *Proceedings of the* 19^{th} *International Workshop on Statistical Modelling*, Biggeri, A., Dreassi, E., Lagazio, C. and Marchi, M. (Eds.). Florence, Italy, p. 465-469.

[89] Nunes, S., Mexia,J.T. and Minder, C. (2004-B). Logit Model for Tuberculosis in Europe (1995-2000). Analysis by Sex and Age Group. In: *Colloquium Biometryczne*, Vol. **34**, p. 147-159.

[90] Oliveira, M.M. e Mexia, J.T. (2004). AIDS in Portugal: endemic versus epidemic forecasting scenarios for mortality. *International Journal of Forecasting* **20**, p. 131-135.

[91] Oliveira, T. (1991). *Probabilidades e Estatística: Conceitos, Métodos e Aplicações* (Vol. II). McGraw-Hill.

[92] Pearl, R. and Reed, L.J. (1920). On the rate of growth of the population of the United States since 1870 and its mathematical representation. *Proceedings of the National Academy of Sciences* **6**, p. 275-288.

[93] Penrose, R.A. (1955). A Generalized Inverse for Matrices. *Proceedings of the Cambridge Philosophical Society* **51**, p. 406-413.

[94] Pereira, D.G. (2004). *Análise Conjunta Pesada de Regressões em Redes de Ensaios*. Dissertação apresentada para obtenção do Grau de Doutor em Matemática na Universidade de Évora.

[95] Pereira, D.G. and Mexia, J.T. (2004). Nodes of the upper contour in Joint Regression Analysis. In: *Colloquium Biometryczne*, Vol. **34**, p. 267-277.

[96] Petrov, V.V. (1962). Vestnik Leningrad. Univ. No. 19, 150-153.

[97] Petrov, V.V. (1975). Sums of independent random variables. *Series Title: Ergebnisse der Mathematik und ihrer Grenzgebiete*, Vol. **82**. Springer-Verlag, Berlin, New York.

[98] Petrov, V.V. (1995). Limit theorems of probability theory: sequences of independent random variables. *Series Title: Oxford studies in probability*, **4**. Clarendon Press, Oxford, Oxford University Press, New York.

[99] Pierce, B.A. (2005). *Genetics: A Conceptual Approach*. 2nd edition. W. H. Freeman and Company.

[100] Pina, H. (1995). *Métodos Numéricos*. McGraw-Hill Portugal.

[101] Popper, K.R. (1995). *The Logic of Scientific Discovery*. Routledge. London and New York.

[102] Reed, L.J. and Berkson, J. (1929). The application of the logistic function to experimental data. *Journal of Physical Chemistry* **33**, p. 760-779.

[103] Rohagti, V.K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. John Wiley and Sons.

[104] Rose, C.and Smith, M.D. (2002). K-Statistics: Unbiased Estimators of Cumulants. *Series 7.2C in Mathematical Statistics with Mathematica*. New York: Springer Verlag, p. 256-259.

[105] Scheffé, H. (1959). *The Analysis of Variance*. John Wiley and Sons.

[106] Seber, G.A.F. (1980). *The Linear Hypothesis: A General Theory*. (2nd Edition). Charles Griffin and Company LTD, London.

[107] Stern, C. and Sherwood, E.R. (1966). *The Origin of Genetics - A Mendel Source Book*. W. H. Freeman, San Francisco.

[108] Strang, G. (1988). *Linear Algebra and its Applications.* (3rd edition). Brooks Cole.

[109] *Surveillance of Tuberculosis in Europe - Euro TB.* Report on Tuberculosis Cases Notified in 1997, 1998, 1999, 2000. Institut de Veille Sanitaire, France. WHO Collaborating Centre for the Surveillance of Tuberculosis in Europe. Royal Netherlands Tuberculosis Association (KNCV): http://www.eurotb.org

[110] Vanpaemel, G. (1987). Quetelet en Verhulst over de mathematische wetten van de bevolkingsgroei. *Academiae Abacleta, Mededelingen van de Koninklijke Academie voor Wetenschappen, Letteren en Schoone Kunsten van België* **49**, p. 99-114.

[111] Verhulst, P.-F. (1838). Notice sur la loi que la population suit dans son accroissement. *Correspondance Mathématique et Physique, publiée par A. Quetelet* **10**, p. 113.

[112] Verhulst, P.-F. (1845). Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique* **18**, p. 1-38.

[113] Verhulst, P.-F. (1847). Deuxième mémoire sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique* **20**, p. 1-32.

[114] Weisstein, E.W. (2002). *CRC Concise Encyclopedia of Mathematics* (2nd edition). Chapman and Hall.

[115] WHO Report - Global Tuberculosis Control (1997-2001). Global Tuberculosis Control. Surveillance, Planning, Financing. Communicable Diseases. World Health Organization. Geneva: http://www.who.int/tb/publications.

[116] Wilson, E.B. and Worcester, J. (1943). The determination of L.D. 50 and its sampling error in bio-assay. *Proceedings of the National Academy of Sciences* **29**, p. 79. First of a series of three articles.

[117] Yule, G.U. (1925). The growth of population and the factors which control it. *Journal of the Royal Statistical Society* **138**, p. 1-59.